A Proofs

In order to mathematically analyze transformation-based self-supervised (or data augmentation) GANs, we need to rewrite the objective functions that are easy to get derivatives. Considering that the data $x \in \mathcal{X}$ and the transformation $T_k \in \mathcal{T}$ are independent from each other and the transformed data $\tilde{x} \in \tilde{\mathcal{X}} = \mathcal{T}(\mathcal{X})$ is deterministic depended on both x and T_k, we have $p_d(\tilde{x}, x, T_k) = p_d(x)p(T_k)p(\tilde{x}|x, T_k) = p_d(x)p(T_k)\delta(\tilde{x} - T_k(x))$ and $p_g(\tilde{x}, x, T_k) = p_g(x)p(T_k)p(\tilde{x}|x, T_k) = p_g(x)p(T_k)\delta(\tilde{x} - T_k(x))$ with the indicator function $\delta(0) = 1$ and $\delta(\tilde{x}) = 0, \forall \tilde{x} \neq 0$.

Proposition 2. *For any continuous and differentiable function f whose domain is $\tilde{\mathcal{X}}$, we have:*

$$\mathbb{E}_{x\sim\mathcal{P},T_k\sim\mathcal{T}}[\log f(T_k(x))] = \mathbb{E}_{\tilde{x}\sim\mathcal{P}^T,T_k\sim\mathcal{T}^{\tilde{x}}}[\log f(\tilde{x})] = \mathbb{E}_{T_k\sim\mathcal{T},\tilde{x}\sim\mathcal{P}^{T_k}}[\log f(\tilde{x})], \tag{12}$$

where $\mathcal{P}$ denotes the original data distribution, $\mathcal{P}^T$ indicates the mixture distribution of transformed data, $\mathcal{P}^{T_k}$ means the distribution of transformed data given the transformation T_k, and $\mathcal{T}^{\tilde{x}}$ represents the distribution of transformation given the transformed data $\tilde{x}$.

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim P,T_k\sim\mathcal{T}}[\log f(T_k(x))] &= \int p(x)\sum_{k=1}^{K} p(T_k)\log f(T_k(x))dx \\
&= \int p(x)\sum_{k=1}^{K} p(T_k)\delta(\tilde{x} - T_k(x))\log f(\tilde{x})dxd\tilde{x} \\
&= \int p(x)\sum_{k=1}^{K} p(T_k)p(\tilde{x}|x, T_k)\log f(\tilde{x})dxd\tilde{x} \\
&= \int \sum_{k=1}^{K} p(\tilde{x}, x, T_k)\log f(\tilde{x})dxd\tilde{x} \\
&= \int \sum_{k=1}^{K} p(\tilde{x}, T_k)\log f(\tilde{x})d\tilde{x} \\
&= \int p^T(\tilde{x})\sum_{k=1}^{K} p_d(T_k|\tilde{x})\log f(\tilde{x})d\tilde{x} \\
&= \mathbb{E}_{\tilde{x}\sim P^T,T_k\sim\mathcal{T}^{\tilde{x}}}[\log f(\tilde{x})] \\
&= \int \sum_{k=1}^{K} p(T_k)p(\tilde{x}|T_k)\log f(\tilde{x})d\tilde{x} \\
&= \mathbb{E}_{T_k\sim\mathcal{T},\tilde{x}\sim\mathcal{P}^{T_k}}[\log f(\tilde{x})].
\end{aligned} \tag{13}$$

□

A.1 Proof of Theorem 1

This Theorem 1 was proved in the SSGAN-MS paper [52]. We here give a brief proof for completeness of this paper. Readers are encouraged to refer to the original proof in [52] for more details.

Theorem 1 ([52])**.** *Given the optimal classifier $C^*(k|\tilde{x}) = \frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^{K} p_d^{T_k}(\tilde{x})}$ of SSGAN, at the equilibrium point, maximizing the self-supervised task for the generator is equivalent to:*

$$\max_G \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^{T_k}}\log\left(\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^{K} p_d^{T_k}(\tilde{x})}\right)\right], \tag{3}$$

where $\mathcal{P}_g^{T_k}$, $\mathcal{P}_d^{T_k}$ indicate the distribution of transformed generated or real data $\tilde{x} \in \tilde{\mathcal{X}}$ under the transformation T_k with density of $p_g^{T_k}(\tilde{x}) = \int \delta(\tilde{x} - T_k(x))p_g(x)dx$ or $p_d^{T_k}(\tilde{x}) = \int \delta(\tilde{x} - T_k(x))p_d(x)dx$.

Proof. According to Proposition 2, the objective function of the self-supervised task for the classifier of SSGAN can be rewritten as follows:

$$\max_C \mathbb{E}_{x\sim\mathcal{P}_d, T_k\sim\mathcal{T}}[\log C(k|T_k(x))] \Rightarrow \max_C \mathbb{E}_{\tilde{x}\sim\mathcal{P}_d^T, T_k\sim\mathcal{T}_d^{\tilde{x}}}[\log C(k|\tilde{x})]. \tag{14}$$

According to the Proposition 1 in [52], the optimal classifier C^* has the form of:

$$C^*(k|\tilde{x}) = \frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}. \tag{15}$$

Therefore, the objective function of the self-supervised task for the generator of SSGAN, under the optimal classifier, can be considered as follows:

$$\begin{aligned} &\max_G \mathbb{E}_{x\sim\mathcal{P}_g, T_k\sim\mathcal{T}}[\log C^*(k|T_k(x))] \\ \Rightarrow &\max_G \mathbb{E}_{T_k\sim\mathcal{T}, \tilde{x}\sim\mathcal{P}_g^{T_k}}[\log C^*(k|\tilde{x})] \\ \Rightarrow &\max_G \frac{1}{K}\sum_{k=1}^K \left[\mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^{T_k}} \log\left(\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}\right)\right]. \end{aligned} \tag{16}$$

□

A.2 Proof of Theorem 2

Theorem 2. *Given the optimal classifier $C_+^*(k|\tilde{x}) = \frac{p_d^T(\tilde{x})}{p_g^T(\tilde{x})}\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}C_+^*(0|\tilde{x})$ of SSGAN-MS, at the equilibrium point, maximizing the self-supervised task for the generator is equivalent to*[2]*:*

$$\min_G \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^T \| \mathcal{P}_d^T) - \frac{1}{K}\sum_{k=1}^K \left[\mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^{T_k}} \log\left(\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}\right)\right], \tag{5}$$

where $\mathcal{P}_g^T$, $\mathcal{P}_d^T$ represent the mixture distribution of transformed generated or real data $\tilde{x} \in \tilde{\mathcal{X}}$ with density of $p_g^T(\tilde{x}) = \sum_{k=1}^K p(T_k) p_g^{T_k}(\tilde{x})$ or $p_d^T(\tilde{x}) = \sum_{k=1}^K p(T_k) p_d^{T_k}(\tilde{x})$.

Proof. According to Proposition 2, we first rewrite the objective function of the self-supervised task for the classifier of SSGAN-MS as follows:

$$\begin{aligned} &\max_{C_+} \mathbb{E}_{x\sim\mathcal{P}_d, T_k\sim\mathcal{T}}[\log C_+(k|T_k(x))] + \mathbb{E}_{x\sim\mathcal{P}_g, T_k\sim\mathcal{T}}[\log C_+(0|T_k(x))] \\ \Rightarrow &\max_{C_+} \mathbb{E}_{\tilde{x}\sim\mathcal{P}_d^T, T_k\sim\mathcal{T}_d^{\tilde{x}}}[\log C_+(k|\tilde{x})] + \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T, T_k\sim\mathcal{T}_g^{\tilde{x}}}[\log C_+(0|\tilde{x})]. \end{aligned} \tag{17}$$

According to the Proposition 2 of [52], for any fixed generator, the optimal classifier C_+^* is:

$$C_+^*(k|\tilde{x}) = \frac{p_d^T(\tilde{x})}{p_g^T(\tilde{x})}\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}C_+^*(0|\tilde{x}), \forall k \in \{1, 2, \cdots, K\}. \tag{18}$$

Since $\sum_{k=0}^K C_+^*(k|\tilde{x}) = 1$ for each transformed data $\tilde{x} \in \tilde{\mathcal{X}}$, we have:

$$\begin{aligned} C_+^*(0|\tilde{x}) &= \frac{p_g^T(\tilde{x})}{p_d^T(\tilde{x}) + p_g^T(\tilde{x})}, \\ C_+^*(k|\tilde{x}) &= \frac{p_d^T(\tilde{x})}{p_d^T(\tilde{x}) + p_g^T(\tilde{x})}\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})} = \frac{p(T_k)p_d^{T_k}(\tilde{x})}{p_d^T(\tilde{x}) + p_g^T(\tilde{x})}, \forall k \in \{1, 2, \cdots, K\}. \end{aligned} \tag{19}$$

[2]Note that our Theorem 2 corrects the wrong version in the SSGAN-MS paper [52], where the authors mistakenly regard $\frac{p_d^T(\tilde{x})}{p_g^T(\tilde{x})} = \frac{\sum_{k=1}^K p(T_k)p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p(T_k)p_g^{T_k}(\tilde{x})}$ as $\frac{p_d^{T_k}(\tilde{x})}{p_g^{T_k}(\tilde{x})}$ in their proof. Please see Appendix A.2 for details.

The self-supervised task for the generator of SSGAN-MS, under the optimal classifier, is equal to:

$$
\begin{aligned}
&\max_G \mathbb{E}_{x\sim\mathcal{P}_g,T_k\sim\mathcal{T}}[\log C_+^*(k|T_k(x))] - \mathbb{E}_{x\sim\mathcal{P}_g,T_k\sim\mathcal{T}}[\log C_+^*(0|T_k(x))] \\
\Rightarrow &\max_G \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T,T_k\sim\mathcal{T}_g^{\tilde{x}}}[\log C_+^*(k|\tilde{x})] - \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T,T_k\sim\mathcal{T}_g^{\tilde{x}}}[\log C_+^*(0|\tilde{x})] \\
\Rightarrow &\max_G \int p_g^T(\tilde{x}) \sum_{k=1}^K p_g(T_k|\tilde{x}) \log\left(\frac{p_d^T(\tilde{x})}{p_d^T(\tilde{x})+p_g^T(\tilde{x})}\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}\right) d\tilde{x} \\
&- \int p_g^T(\tilde{x}) \sum_{k=1}^K p_g(T_k|\tilde{x}) \log\left(\frac{p_g^T(\tilde{x})}{p_d^T(\tilde{x})+p_g^T(\tilde{x})}\right) d\tilde{x} \\
\Rightarrow &\max_G \int p_g^T(\tilde{x}) \sum_{k=1}^K p_g(T_k|\tilde{x}) \log\left(\frac{p_d^T(\tilde{x})}{p_g^T(\tilde{x})}\right) d\tilde{x} + \int p_g^T(\tilde{x}) \sum_{k=1}^K p_g(T_K|\tilde{x}) \log\left(\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}\right) d\tilde{x} \\
\Rightarrow &\max_G \int p_g^T(\tilde{x}) \log\left(\frac{p_d^T(\tilde{x})}{p_g^T(\tilde{x})}\right) d\tilde{x} + \sum_{k=1}^K p(T_K) \int p_g^{T_k}(\tilde{x}) \log\left(\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}\right) d\tilde{x} \\
\Rightarrow &\min_G \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^T \| \mathcal{P}_d^T) - \frac{1}{K}\sum_{k=1}^K \left[\mathbb{E}_{\tilde{x}\sim P_g^{T_k}} \log\left(\frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K p_d^{T_k}(\tilde{x})}\right)\right].
\end{aligned}
\tag{20}
$$

□

A.3 Proof of Proposition 1

Proposition 1. *For any fixed generator, given a data $\tilde{x} \in \tilde{\mathcal{X}}$ that drawn from mixture distribution of transformed data, the optimal label-augmented discriminator of SSGAN-LA has the form of:*

$$
D_{\mathrm{LA}}^*(k,1|\tilde{x}) = \frac{p_d^{T_k}(\tilde{x})}{\sum_{k=1}^K (p_d^{T_k}(\tilde{x}) + p_g^{T_k}(\tilde{x}))}, D_{\mathrm{LA}}^*(k,0|\tilde{x}) = \frac{p_g^{T_k}(\tilde{x})}{\sum_{k=1}^K (p_d^{T_k}(\tilde{x}) + p_g^{T_k}(\tilde{x}))}. \tag{7}
$$

Proof. We can first rewrite the objective function for the label-augmented discriminator as follows:

$$
\begin{aligned}
&\max_{D_{\mathrm{LA}}} \mathbb{E}_{x\sim\mathcal{P}_d,T_k\sim\mathcal{T}}[\log D_{\mathrm{LA}}(k,1|T_k(x))] + \mathbb{E}_{x\sim\mathcal{P}_g,T_k\sim\mathcal{T}}[\log D_{\mathrm{LA}}(k,0|T_k(x))] \\
\Rightarrow &\max_{D_{\mathrm{LA}}} \mathbb{E}_{\tilde{x}\sim\mathcal{P}_d^T,T_k\sim\mathcal{T}_d^{\tilde{x}}}[\log D_{\mathrm{LA}}(k,1|\tilde{x})] + \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T,T_k\sim\mathcal{T}_g^{\tilde{x}}}[\log D_{\mathrm{LA}}(k,0|\tilde{x})] \\
\Rightarrow &\max_{D_{\mathrm{LA}}} \int p_d^T(\tilde{x}) \sum_{k=1}^K p_d(T_k|\tilde{x}) \log D_{\mathrm{LA}}(k,1|\tilde{x}) d\tilde{x} + \int p_g^T(\tilde{x}) \sum_{k=1}^K p_g(T_k|\tilde{x}) \log D_{\mathrm{LA}}(k,0|\tilde{x}) d\tilde{x}.
\end{aligned}
\tag{21}
$$

Maximizing this integral is equivalent to maximize the component for every transformed data $\tilde{x} \in \tilde{\mathcal{X}}$:

$$
\begin{aligned}
&\max_{D_{\mathrm{LA}}} p_d^T(\tilde{x}) \sum_{k=1}^K p_d(T_k|\tilde{x}) \log D_{\mathrm{LA}}(k,1|\tilde{x}) + p_g^T(\tilde{x}) \sum_{k=1}^K p_g(T_k|\tilde{x}) \log D_{\mathrm{LA}}(k,0|\tilde{x}), \\
&\text{s.t.} \sum_{k=1}^K D_{\mathrm{LA}}(k,1|\tilde{x}) + D_{\mathrm{LA}}(k,0|\tilde{x}) = 1.
\end{aligned}
\tag{22}
$$

Define the Lagrange function as follows:

$$
\mathcal{L}(D_{\mathrm{LA}}, \lambda) = \sum_{i=0}^1 p_i^T(\tilde{x}) \sum_{k=1}^K p_i(T_k|\tilde{x}) \log D_{\mathrm{LA}}(k,i|\tilde{x}) + \lambda(\sum_{i=0}^1 \sum_{k=1}^K D_{\mathrm{LA}}(k,i|\tilde{x}) - 1), \tag{23}
$$

with $p_1^T(\tilde{x}) = p_d^T(\tilde{x})$, $p_0^T(\tilde{x}) = p_g^T(\tilde{x})$, $p_1(T_k|\tilde{x}) = p_d(T_k|\tilde{x})$ and $p_0(T_k|\tilde{x}) = p_g(T_k|\tilde{x})$, and $\lambda \in \mathbb{R}$ the Lagrange multiplier.

Get the derivatives with respect to $D_{\mathrm{LA}}(k, i|\tilde{x})$ and λ, and then let them equals to 0, then we have:

$$\frac{\partial \mathcal{L}}{\partial D_{\mathrm{LA}}(k,i|\tilde{x})} = \frac{p_i^T(\tilde{x}) p_i(T_k|\tilde{x})}{D_{\mathrm{LA}}(k,i|\tilde{x})} + \lambda = 0, \frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{i=0}^{1}\sum_{k=1}^{K} D_{\mathrm{LA}}(k,i|\tilde{x}) - 1 = 0. \tag{24}$$

By solving the above equations and according to $\frac{\partial^2 \mathcal{L}}{\partial D_{\mathrm{LA}}(k,i|\tilde{x})^2} < 0$, we obtain the optimal label-augmented discriminator for the transformed data $\tilde{x}$ as follows:

$$D^*_{\mathrm{LA}}(k,i|\tilde{x}) = \frac{p_i^T(\tilde{x}) p_i(T_k|\tilde{x})}{\sum_{i=0}^{1}\sum_{k=1}^{K} p_i^T(\tilde{x}) p_i(T_k|\tilde{x})} = \frac{p(T_k) p_i(\tilde{x}|T_k)}{\sum_{i=0}^{1}\sum_{k=1}^{K} p(T_k) p_i(\tilde{x}|T_k)} = \frac{p_i(\tilde{x}|T_k)}{\sum_{i=0}^{1}\sum_{k=1}^{K} p_i(\tilde{x}|T_k)}. \tag{25}$$

The third equation holds because of $p(T_k) = \frac{1}{K}, \forall T_k \in \mathcal{T}$. This concludes the proof as the defined notations: $p_1(\tilde{x}|T_k) = p_d(\tilde{x}|T_k) = p_d^{T_k}(\tilde{x})$ and $p_0(\tilde{x}|T_k) = p_g(\tilde{x}|T_k) = p_g^{T_k}(\tilde{x})$.

□

A.4 Proof of Theorem 3

Theorem 3. *The objective function for the generator of SSGAN-LA, given the label-augmented optimal discriminator, boils down to:*

$$\min_G \frac{1}{K}\sum_{k=1}^{K} \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^{T_k} \| \mathcal{P}_d^{T_k}). \tag{9}$$

The global minimum is achieved if and only if $\mathcal{P}_g = \mathcal{P}_d$ when $\exists T_k \in \mathcal{T}$ is an invertible transformation.

Proof.

$$\begin{aligned}
&\max_G \mathbb{E}_{x\sim\mathcal{P}_g, T_k\sim\mathcal{T}}[\log D^*_{\mathrm{LA}}(k,1|T_k(x))] - \mathbb{E}_{x\sim\mathcal{P}_g}\mathbb{E}_{T_k\sim\mathcal{T}}[\log D^*_{\mathrm{LA}}(k,0|T_k(x))] \\
\Rightarrow &\max_G \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T, T_k\sim\mathcal{T}_g^{\tilde{x}}}[\log D^*_{\mathrm{LA}}(k,1|\tilde{x})] - \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T, T_k\sim\mathcal{T}_g^{\tilde{x}}}[\log D^*_{\mathrm{LA}}(k,0|\tilde{x})] \\
\Rightarrow &\max_G \int p_g^T(\tilde{x}) \sum_{k=1}^{K} p_g(T_k|\tilde{x}) \log D^*_{\mathrm{LA}}(k,1|\tilde{x}) d\tilde{x} \\
&- \int p_g^T(\tilde{x}) \sum_{k=1}^{K} p_g(T_k|\tilde{x}) \log D^*_{\mathrm{LA}}(k,0|\tilde{x}) d\tilde{x} \\
\Rightarrow &\max_G \int p_g^T(\tilde{x}) \sum_{k=1}^{K} p_g(T_k|\tilde{x}) \log \frac{p_d(\tilde{x}|T_k)}{\sum_{k=1}^{K}(p_d(\tilde{x}|T_k) + p_g(\tilde{x}|T_k))} d\tilde{x} \\
&- \int p_g^T(\tilde{x}) \sum_{k=1}^{K} p_g(T_k|\tilde{x}) \log \frac{p_g(\tilde{x}|T_k)}{\sum_{k=1}^{K}(p_d(\tilde{x}|T_k) + p_g(\tilde{x}|T_k))} d\tilde{x} \\
\Rightarrow &\max_G \int p_g^T(\tilde{x}) \sum_{k=1}^{K} p_g(T_k|\tilde{x}) \log \frac{p_d(\tilde{x}|T_k)}{p_g(\tilde{x}|T_k)} d\tilde{x} \\
\Rightarrow &\max_G \sum_{k=1}^{K} p(T_k) \int p_g(\tilde{x}|T_k) \log \frac{p_d(\tilde{x}|T_k)}{p_g(\tilde{x}|T_k)} d\tilde{x} \\
\Rightarrow &\min_G \frac{1}{K}\sum_{k=1}^{K} \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^{T_k} \| \mathcal{P}_d^{T_k}).
\end{aligned} \tag{26}$$

According to Theorem 3 in Appendix of [52], the KL divergence is invariant to invertible/affine transformation. In other words, $\mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^{T_k}\|\mathcal{P}_d^{T_k}) = \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g\|\mathcal{P}_d)$ when the transformation T_k is invertible, which induces $\mathcal{P}_g = \arg\min_G \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^{T_k}\|\mathcal{P}_d^{T_k})\#\mathcal{P}_z = \arg\min_G \mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g\|\mathcal{P}_d)\#\mathcal{P}_z = \mathcal{P}_d$. In addition, $\mathcal{P}_g = \mathcal{P}_d$ is a sufficient condition for $\mathcal{P}_g^{T_{k'}} = \mathcal{P}_d^{T_{k'}}$ that fully minimizes $\mathbb{D}_{\mathrm{KL}}(\mathcal{P}_g^{T_{k'}}\|\mathcal{P}_d^{T_{k'}})$ regardless of the invertibility $T_{k'}$. Therefore, the global maximum of the objective function for the generator of SSGAN-LA given the optimal label-augmented discriminator is achieved if and only if $\mathcal{P}_g = \mathcal{P}_d$ when existing a transformation is invertible.

□

A.5 Proof of Theorem 4

Theorem 4. *At the equilibrium point of DAGAN, the optimal generator implies $\mathcal{P}_g^T = \mathcal{P}_d^T$. However, if $(\mathcal{T}, \circ)$ forms a group and $T_k \in \mathcal{T}$ is uniformly sampled, then the probability that the optimal generator replicates the real data distribution is $\mathbb{P}(\mathcal{P}_g = \mathcal{P}_d|\mathcal{P}_g^T = \mathcal{P}_d^T) = 0$.*

Proof. We first prove the first sentence in this Theorem. According to Proposition 2, the objective function of DAGAN can be rewritten as follows:

$$\begin{aligned}\min_G \max_D V(G,D) =& \mathbb{E}_{x\sim\mathcal{P}_d, T_k\in\mathcal{T}}[\log D(T_k(x))] + \mathbb{E}_{x\sim\mathcal{P}_g, T_k\in\mathcal{T}}[\log(1-D(T_k(x)))] \\ =& \mathbb{E}_{\tilde{x}\sim\mathcal{P}_d^T, T_k\in\mathcal{T}_d^{\tilde{x}}}[\log D(\tilde{x})] + \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T, T_k\in\mathcal{T}_g^{\tilde{x}}}[\log(1-D(\tilde{x}))] \\ =& \mathbb{E}_{\tilde{x}\sim\mathcal{P}_d^T}[\log D(\tilde{x})] + \mathbb{E}_{\tilde{x}\sim\mathcal{P}_g^T}[\log(1-D(\tilde{x}))].\end{aligned} \tag{27}$$

According to the Theorem 1 in [13], the global minimum of the virtual training criterion $V(G, D^*)$, given the optimal discriminator D^*, is achieved if and only if $\mathcal{P}_g^T = \mathcal{P}_d^T$.

We then prove the second sentence in this Theorem. The main idea of the proof is to construct countless generated distributions who satisfy the equilibrium point of DAGAN. However, there is only one real data distribution. Therefore, the probability that the generator of DAGAN learns the real data distribution is 0 even though at its equilibrium point.

Since the set of transformations $\mathcal{T}$ forms a group with respect to the composition operator $\circ$, and according to the **closure property** of group, the composition of any two transformations is also in the set (i.e., $T_i \circ T_j \in \mathcal{T}, \forall T_i, T_j \in \mathcal{T}$). In addition, according to the converse-negative proposition of the **cancellation law** of group, the compositions of a transformation with other different transformations are different from each other (i.e., $T_i \circ T_k \neq T_j \circ T_k, \forall T_i \neq T_j, T_k \in \mathcal{T}$). Based on the above properties and **inclusion–exclusion principle**, we have $\{T_j \circ T_i | T_i \in \mathcal{T}\} = \mathcal{T}, \forall T_j \in \mathcal{T}$.

Let us construct a family of distribution $\mathcal{P}_\pi$ with density of $p_\pi(\hat{x}) = \sum_{j=1}^K \pi_j p_d(\hat{x}|T_j) = \sum_{j=1}^K \pi_j \int p_d(x)p(\hat{x}|x, T_j)dx$ with mixture weights $\Pi = \{\pi_j\}_{j=1}^K$, subject to $\sum_{j=1}^K \pi_j = 1$ and

$0 \leq \pi_j \leq 1, \forall \pi_j \in \Pi$, then the mixture transformed distribution of data from $\mathcal{P}_\pi$ is:

$$
\begin{aligned}
p_\pi^T(\tilde{x}) &= \int p_\pi(\hat{x}) \sum_{i=1}^{K} p(T_i) p(\tilde{x}|\hat{x}, T_i) d\hat{x} \\
&= \int \sum_{j=1}^{K} \pi_j \int p_d(x) p(\hat{x}|x, T_j) dx \sum_{i=1}^{K} p(T_i) p(\tilde{x}|\hat{x}, T_i) d\hat{x} \\
&= \sum_{j=1}^{K} \pi_j \int \sum_{i=1}^{K} p_d(x) p(\hat{x}|x, T_j) p(T_i) p(\tilde{x}|\hat{x}, T_i) dx d\hat{x} \\
&\overset{\text{ⓐ}}{=\!=} \sum_{j=1}^{K} \pi_j \int \sum_{i=1}^{K} p(T_i) p_d(x) p(\tilde{x}, \hat{x}|x, T_i, T_j) dx d\hat{x} \\
&= \sum_{j=1}^{K} \pi_j \int \sum_{i=1}^{K} p(T_i) p_d(x) p(\tilde{x}|x, T_i, T_j) dx \\
&= \sum_{j=1}^{K} \pi_j \int \sum_{i=1}^{K} \frac{1}{K} p_d(x) p(\tilde{x}|x, T_i, T_j) dx \\
&\overset{\text{ⓑ}}{=\!=} \sum_{j=1}^{K} \pi_j \int \sum_{k=1}^{K} \frac{1}{K} p_d(x) p(\tilde{x}|x, T_k) dx \\
&= \sum_{j=1}^{K} \pi_j \int \sum_{k=1}^{K} p(T_k) p_d(x) p(\tilde{x}|x, T_k) dx \\
&= \sum_{j=1}^{K} \pi_j p_d^T(\tilde{x}) \\
&= p_d^T(\tilde{x}).
\end{aligned}
\tag{28}
$$

The equation ⓐ holds because of $p(\hat{x}|x, T_j) = p(\hat{x}|x, T_i, T_j)$ ($\hat{x}$ only depends on x and T_j) and $p(\tilde{x}|\hat{x}, T_i) = p(y|\hat{x}, x, T_i, T_j)$ ($\tilde{x}$ is independent of x and T_j given $\hat{x}$ and T_i). The equation ⓑ holds because of $p(\tilde{x}|x, T_i, T_j) = \delta(\tilde{x} - T_j \circ T_i(x))$ and $\{T_j \circ T_i | T_i \in \mathcal{T}\} = \mathcal{T}, \forall T_j \in \mathcal{T}$.

Therefore, there are infinite generator distributions $\mathcal{P}_\pi$ that satisfy the equilibrium point of DAGAN ($\mathcal{P}_\pi^T = \mathcal{P}_d^T$), but only one of them is the desired generator distribution (i.e., the real data distribution $\mathcal{P}_d$). In particular, we have $\mathcal{P}_\pi = \mathcal{P}_d$ if and only if $\pi_1 = 1, \pi_k = 0, \forall k = \{2, 3, \cdots, K\}$, and the corresponding probability is $\mathbb{P}(\mathcal{P}_\pi = \mathcal{P}_d | \mathcal{P}_\pi^T = \mathcal{P}_d^T) = \mathbb{P}(\pi_1 = 1, \pi_k = 0, \forall \pi_k \in \Pi \setminus \pi_1 | \sum_{k=1}^{K} \pi_k = 1, 0 \leq \pi_k \leq 1, \forall \pi_k \in \Pi) = 0$. This concludes our proof. □

B Experimental Settings in Section 5.3

We implement all methods based on unconditional BigGAN [4] without the annotated labels. To implement the unconditional BigGAN, we replace the conditional batch normalization in the generator with standard batch normalization and remove the label projection technique in the discriminator. The network structure of generators of all methods is the same, and the network structure of discriminators of all methods is different only in the output layer. We train all methods for 100 epochs with a batch size of 100 on all datasets. The optimizer is Adam with betas $(\beta_1, \beta_2) = (0.0, 0.999)$ for both the generator and discriminator. The learning rate for the generator is 2×10^{-4} on CIFAR-10 and STL-10, and 1×10^{-4} on Tiny-ImageNet, and the learning rate for the discriminator/classifier is 2×10^{-4} on CIFAR-10 and STL-10, and 4×10^{-4} on Tiny-ImageNet. All baselines use the hinge loss [29, 51] as the implementation of the original GAN loss.

$$\begin{aligned} &\min_D V_{\text{H}}^D(G, D) = \mathbb{E}_{x\sim\mathcal{P}_d}[\max(0, 1 - D(x))] + \mathbb{E}_{x\sim\mathcal{P}_g}[\max(0, 1 + D(x))], \\ &\min_G V_{\text{H}}^G(G, D) = \mathbb{E}_{x\sim\mathcal{P}_g}[-D(x)]. \end{aligned} \tag{29}$$

And the proposed SSGAN-LA adopts the multi-class hinge loss as defined in the following:

$$\begin{aligned} &\min_{D_{\text{LA}}} V_{\text{MH}}^{D_{\text{LA}}}(G, D_{\text{LA}}) \\ &\quad = \mathbb{E}_{x\sim\mathcal{P}_d, T_k\sim\mathcal{T}}[\mathbb{E}_{(k',l)\neq(k,1)}[\max(0, 1 - D_{\text{LA}}(k, 1|T_k(x)) + D_{\text{LA}}(k', l|T_k(x)))]] \\ &\quad + \mathbb{E}_{x\sim\mathcal{P}_g, T_k\sim\mathcal{T}}[\mathbb{E}_{(k',l)\neq(k,0)}[\max(0, 1 - D_{\text{LA}}(k, 0|T_k(x)) + D_{\text{LA}}(k', l|T_k(x)))]], \\ &\min_G V_{\text{MH}}^G(G, D_{\text{LA}}) \\ &\quad = \mathbb{E}_{x\sim\mathcal{P}_g, T_k\sim\mathcal{T}}[\mathbb{E}_{(k',l)\neq(k,1)}[-D_{\text{LA}}(k, 1|T_k(x)) + D_{\text{LA}}(k', l|T_k(x)))]] \\ &\quad - \mathbb{E}_{x\sim\mathcal{P}_g, T_k\sim\mathcal{T}}[\mathbb{E}_{(k',l)\neq(k,0)}[-D_{\text{LA}}(k, 0|T_k(x)) + D_{\text{LA}}(k', l|T_k(x)))]]. \end{aligned} \tag{30}$$

The multi-class hinge loss functions are the extended version of the (binary) hinge loss functions. Notice that the discriminators (including the label-augmented discriminator D_{LA}) in the hinge loss functions are no longer required to output a (softmax) probability. Other hyper-parameters in baselines are the same as the authors' suggestions in their papers unless otherwise specified. To obtain the optimal label-augmented discriminator of SSGAN-LA as much as possible and because that the discriminator solves a more challenging classification task, we set the discriminator updating steps per generator step as $n_{\text{dis}} = 4$ for SSGAN-LA. We also set $n_{\text{dis}} = 4$ for GAN, DAGAN, and DAGAN-MD as it performs better in our experiments. We follow the practices in [12, 26, 54] to perform all transformations on each sample for DAGAN, DAGAN-MD, and SSGAN-LA.

C Performance of SSGAN-LA Retaining the Original Discriminator

We investigate the performance of the proposed label-augmented discriminator D_{LA} combined with the original discriminator D under the same data transformation setting as SSGAN [6] that rotate a quarter images in a batch in all four considered directions. Specifically, the objective functions for the original discriminator D, the label-augmented discriminator D_{LA}, and the generator G of the original discriminator retained SSGAN-LA (SSGAN-LA$^+$) are given by:

$$\begin{aligned} &\min_{D, D_{\text{LA}}} V_{\text{H}}^D(G, D) + \lambda_d \cdot V_{\text{MH}}^{D_{\text{LA}}}(G, D_{\text{LA}}), \\ &\min_G V_{\text{H}}^G(G, D) + \lambda_g \cdot V_{\text{MH}}^G(G, D_{\text{LA}}), \end{aligned} \tag{31}$$

where λ_d and λ_g are two hyper-parameters that trade-off the GAN task and the self-supervised task. The discriminator update steps are 2 per generator update step for all methods following the practice of SSGAN. The hyper-parameters λ_d and λ_g of SSGAN, SSGAN-MS, and SSGAN-LA$^+$ are selected from $\{0.2, 1.0\}$ according to the best FID score. Table 4 shows that, under the same data transformation settings as that in the SSGAN paper, SSGAN-LA$^+$ outperforms all competing baselines in terms of FID and IS results, verifying the effectiveness of the proposed self-supervised approach for training GANs.

Table 4: FID (↓) and IS (↑) comparison on CIFAR-10, STL-10, and Tiny-ImageNet. SSGAN-LA$^+$ retains the original discriminator. We use the same data transformation setting of SSGAN.

Dataset	Metric	GAN	SSGAN	SSGAN-MS	DAGAN$^+$	DAGAN-MD	SSGAN-LA$^+$
CIFAR-10	FID	10.83	7.52	7.08	11.07	10.05	**6.64**
	IS	8.42	8.29	8.45	8.10	8.14	**8.51**
STL-10	FID	20.15	16.84	16.46	18.97	21.68	**15.91**
	IS	10.25	10.36	10.40	10.12	10.29	**10.87**
Tiny-ImageNet	FID	31.01	30.09	25.76	58.91	50.14	**24.23**
	IS	9.80	10.21	10.57	7.06	7.80	**10.86**

D Ablation Study on Self-Supervised Task for the Generator of SSGAN

In this experiment, we investigate the effects of the original self-supervised task proposed in [6] for the generator. We change the value of λ_g of SSGAN while keeping $\lambda_d = 1.0$ fixed. As reported in Table 5, SSGAN with $\lambda_g = 0.2$ generally performs better than that with $\lambda_g = 0.0$, showing that the self-supervised task for the generator could benefit the generation performance of the generator. We argue that the reason is that the self-supervised task for the generator can reduce the difficulty of optimizing the generator by providing extra guidance. However, as λ_g increases to 1.0, the self-supervised task for the generator has a negative effect on the generation performance because the implied goal is essentially inconsistent with generative modeling.

Table 5: Ablation study on λ_g of SSGAN.

Dataset	Metric	$\lambda_g = 0.0$	$\lambda_g = 0.2$	$\lambda_g = 1.0$
CIFAR-10	FID	8.07	7.52	8.54
	IS	8.21	8.29	8.41
STL-10	FID	18.04	16.84	18.88
	IS	10.20	10.36	10.03
Tiny-ImageNet	FID	30.69	30.09	30.27
	IS	10.67	10.21	10.23