1. Optimal Transport Regularization

1.1. Proof of the Existence of $y_{\sigma(j)}$

In this section, first we prove that for every $x_j$, we can always find a $y_{\sigma(j)}$ such that $H_{\sigma(j)}^* - H_j^* = c(x_j, y_{\sigma(j)})$ in Section 3.2.

**Proof.** We first prove that if $H_i^*$ and $H_j^*$ are the optimal solutions to the following problem:

$$
\max_{H_i, H_j} \left\{ \frac{1}{m} \sum_{i \in I} H_i - \frac{1}{n} \sum_{j \in J} H_j \right\}
$$

s.t. $H_i - H_j \leq c(x_j, y_i), \quad \forall i \in I, \forall j \in J$

(1)

where $I$ and $J$ are disjoint sets, then for each $x_j$, there exists a $t \in I$, such that $H_i^* - H_j^* = c(x_j, y_t)$.

We prove this by contradiction, i.e., there exists one $x_s$, $s \in J$, such that we cannot find a $y_i$ such that $H_i^* - H_j^* = c(x_s, y_i), \forall i \in I$. This means that $H_s^* > \sup_{i \in I} \{H_i^* - c(x_s, y_i)\}$. We can construct $H'$, such that $H_i' = H_i^*, \forall i \in I, H_j' = H_j^*, \forall j \in J - \{s\}$, and $H_s' = \sup_{i \in I} \{H_i^* - c(x_s, y_i)\}$. It can be verified that $H'$ satisfies the constraints in Eq. (1), and the objective of Eq. (1) achieved at $H'$ is greater than that achieved at $H^*$. This contradicts that $H^*$ is the optimal solution to Eq. (1). Therefore, for every $x_j$, we can find a $y_t, t \in I$, such that $H_i^* - H_j^* = c(x_j, y_t)$.

Next, we will show that if $H_i^* - H_j^* = c(x_j, y_t)$, then $t = \sigma(j)$. Since $c(x_j, y_t) = \frac{K}{2} \|x_j - y_t\|^2$, we have $\frac{K}{2} \|x_j - y_t\|^2 + H_j^* - H_i^* = 0$. According to the definition of $\sigma(j)$ in Eq. (6), and the fact that $H_i^* - H_j^* \leq c(x_j, y_t)$, we have $t = \sigma(j)$.

\hfill $\Box$

1.2. Proof of Eq. (8)

We will show that Theorem 1.17 in Santambrogio’s optimal transport [8], which is based on Brenier’s theorem [11], can also be used to prove Eq. (8). First we will introduce the following theorem.

**Theorem 1.17 in [8]** Given $\nu$ and $\mu$ on a compact domain $\Omega \subset \mathbb{R}^d$, there exists an optimal transport plan $\gamma$ for the cost $c(x, y) = h(x - y)$ with $h$ being strictly convex. Provided $\nu$ is absolutely continuous and $\partial \Omega$ is negligible, there exists a Kantorovich potential $D$. The transport map $T$ can be represented as

$$
T(x) = x + (\nabla h)^{-1}(\nabla D(x))
$$

(2)

where $(\nabla h)^{-1}(\cdot)$ denotes the inverse mapping of $\nabla h(\cdot)$.

Next, we show that the above theorem can also be used to prove Eq. (8).

**Proof.** According to the definition of $\sigma$ in Eq. (6), we know that $x_j$ is transported to $y_{\sigma(j)}$ [9], i.e., $T(x_j) = y_{\sigma(j)}$. If $c(x, y) = \frac{K}{2} \|x - y\|^2$, then

$$
T(x) = x + \frac{1}{K} \nabla D(x)
$$

(3)
For every \( x_j \), we have
\[
T(x_j) = y_{\sigma(j)} = x_j + \frac{1}{K} \nabla D(x_j) \tag{4}
\]
and hence
\[
\nabla D(x_j) + K(x_j - y_{\sigma(j)}) = 0 \tag{5}
\]
This proves Eq. (8).

From the above analysis we can see that Theorem 1.17 in [8] can also lead to the same result shown in Eq. (8).

2. Proofs of Lemmas and Theorems

**Theorem 1** in Section 3

If the discriminator in Eq. (10) has sufficient capacity such that the optimal objective of Eq. (10) is 0, then for any \( \gamma > 0 \), and any optimal solution \( D_w^* \) to Eq. (10),
\[
\frac{1}{m} \sum_{i \in I} D_w^*(y_i) - \frac{1}{n} \sum_{j \in J} D_w^*(x_j) \tag{6}
\]
is the quadratic Wasserstein distance between \( \hat{X} \) and \( \hat{Y} \).

**Proof.** If \( D_w^* \) is an optimizer to Eq. (10), and \( D_w \) has sufficient capacity, then all the three terms in Eq. (10) are 0s when optimal, i.e., \( \frac{1}{m} \sum_{i \in I} D_w^*(y_i) = \frac{1}{m} \sum_{i \in I} H_i^* \), and \( D_w^*(x_j) = H_j^*, \forall j \in J \). Since \( \frac{1}{m} \sum_{i \in I} H_i^* - \frac{1}{n} \sum_{j \in J} H_j^* \) is the quadratic Wasserstein distance between \( \hat{X} \) and \( \hat{Y} \), \( \frac{1}{m} \sum_{i \in I} D_w^*(y_i) - \frac{1}{n} \sum_{j \in J} D_w^*(x_j) \) is the quadratic Wasserstein distance between \( \hat{X} \) and \( \hat{Y} \). \( \square \)

**Lemma 1** in Section 4

The second order derivative of the regularization term
\[
\mathcal{L}_R = \frac{\lambda}{2} \mathbb{E}_{P_r(x)} \left[ \| \nabla_x D_w(x) \|^2 - \| y_{T(x)} - x \|^2 \right] \tag{7}
\]
with respect to \((w, \theta)\) at the equilibrium point is given by:
\[
M_R = \lambda \cdot \mathbb{E}_{P_r(x)} \left[ \nabla_{w,x} D_w^*(x) \nabla_{w,x} D_w^*(x)^\top \right] \tag{8}
\]

**Proof.** At the equilibrium point \((w^*, \theta^*)\), every point \( x \) is transported to \( y_{T(x)} \), and \( x = y_{T(x)} \). Therefore, at the equilibrium point, the regularization term is
\[
\mathcal{L}'_R = \frac{\lambda}{2} \mathbb{E}_{P_r(x)} \left[ \| \nabla_x D_w(x) \|^2 \right] \tag{9}
\]
The first order derivative at the equilibrium point is \( \nabla_w \mathcal{L}'_R = \lambda \cdot \mathbb{E}_{P_r(x)} \left[ \nabla_{w,x} D_w^*(x) \nabla_{w,x} D_w^*(x)^\top \right] \). By applying the chain rule to \( \nabla_w \mathcal{L}'_R \) and use \( \nabla_x D_w^*(x) = 0 \) \((r(w^*) = 0)\) in Eq. (18) and \( P_r(x) = P_r(y) \) at the equilibrium point, we have
\[
\lambda \cdot \mathbb{E}_{P_r(x)} \left[ \nabla_{w,x} D_w^*(x) \nabla_{w,x} D_w^*(x)^\top \right]
\]
\( \square \)

**Lemma 2** in Section 4

The Jacobian of the gradient field \( g(w, \theta) \) at the equilibrium point \((w^*, \theta^*)\) is given by:
\[
g' (w^*, \theta^*) = \begin{pmatrix} M_{DD} + M_R & M_{GD} \\ 0 & 0 \end{pmatrix} \tag{10}
\]
where \( M_R \) is defined in Lemma 1,
\[
M_{DD} = \mathbb{E}_{P_r(y)} \left[ \nabla_w D_w^*(y) \right] \cdot \mathbb{E}_{P_r(y)} \left[ \nabla_w D_w^*(y)^\top \right] + \mathbb{E}_{P_r(y)} \left[ \nabla_y D_w^*(y) \nabla_y D_w^*(y)^\top \right] \tag{11}
\]
\[
M_{GD} = -\mathbb{E}_{P_r(x)} \left[ \sum_{i \in I} D_w^*(x_i) \nabla_{\theta \theta} G_{y^*} (z) \right] \tag{12}
\]
and \( M_{DD} + M_R \) is positive definite.
Proof. We give the first order derivatives of $\mathcal{L}_D$ and $\mathcal{L}_G$ below:

$$\nabla_w \mathcal{L}_D(w, \theta) = (\mathbb{E}_{P_r(y)} [D_w(y)] - \mathbb{E}_{P_r(y)} [H_r(y)]) \cdot \mathbb{E}_{P_r(y)} [\nabla_w D_w(y)] + \mathbb{E}_{P_r(y)} [(D_w(x) - H_s(x)) \nabla_w D_w(x)] + \nabla_w \mathcal{L}_R$$

(13)

$$\nabla_\theta \mathcal{L}_G(w, \theta) = -\mathbb{E}_{P(z)} [\nabla_\theta G_\theta(z) \nabla_z D_w(x)]$$

(14)

Then, we apply the chain rule to Eqs. (13) and (14). We give the second order derivatives below:

$$\nabla^2_{w, \theta} \mathcal{L}_D(w, \theta) = \mathbb{E}_{P_r(x)} [(D_w(x) - H_s(x)) \nabla_\theta G_\theta(z) \nabla^2_w D_w(x)] + \mathbb{E}_{P_r(x)} [\nabla_\theta G_\theta(z) \nabla_D D_w(x) \nabla_w D_w(x)]$$

(15)

$$\nabla^2_\theta \mathcal{L}_G(w, \theta) = -\mathbb{E}_{P(z)} [\nabla_\theta G_\theta(z) \nabla^2_D D_w(x) \nabla_\theta G_\theta(z)] - \mathbb{E}_{P(z)} [\nabla^2_\theta G_\theta(z) \nabla_D D_w(x)]$$

(16)

Since at equilibrium, $D_w^*(x) = D_w^*(y) = 0$, $\nabla_D D_w^*(x) = 0$, $D_w^*(y) = H_s(x)$, $x = y_T(x)$ and $\nabla_D^2 R = M_R$, we have $\nabla^2_D \mathcal{L}(w^*, \theta^*) = M_{DD} + M_R$, $\nabla^2_{\theta, \theta} \mathcal{L}(w^*, \theta^*) = 0$, $\nabla^2_{w, \theta} \mathcal{L}(w^*, \theta^*) = 0$, and $\nabla^2_\theta \mathcal{L}(w^*, \theta^*) = 0$.

By applying Assumption II in our paper and the reparametrization technique in [7], the proof that $M_R$ is positive definite is identical to the proof that $L_{DD}$ is positive definite in Lemma D.5 in [7]. As $M_{DD}$ is non-negative definite, $M_{DD} + M_R$ is positive definite.

Lemma 3 in Section 4

For simultaneous gradient updates of $(w, \theta)$ in WGAN-QC using Eq. (15), if $w = w^*$, then $\theta = \theta^*$.

Proof. First we show that $||\nabla_x D_w^*(x)|| = 0, \forall x \sim P_s(x)$. According to the definition of $w^*$ in Eq. (17), we can find a $\theta^*$, such that $\mathcal{L}_D(w^*, \theta^*) = 0$. According to the definition of $\theta^*$ in Eq. (17), we know that $||y_T(x) - x|| = 0, \forall x \sim P_s(x)$. Therefore, $\mathbb{E}_{P_r(x)} [||\nabla_x D_w(x)||] = 0$, and hence $||\nabla_x D_w^*(x)|| = 0, \forall x \sim P_s(x)$.

Since all the eigenvalues of the Jacobian of the gradient field $g(w, \theta)$ are non-negative, and the positive ones are the same as the eigenvalues of the Hessian of the discriminator loss in Eq. (12), when $w$ reaches $w^*$ of an equilibrium point in WGAN-QC using simultaneous gradient update (Eq. (15)), $w$ is also an optimal solution to Eq. (12) for some $\theta$. Thus, $\mathcal{L}_D(w^*, \theta) = 0$, and we have

$$\mathbb{E}_{P_r(x)} [||\nabla_x D_w^*(x)|| - K||y_T(x) - x||^2] = 0$$

(19)

Recall that $||\nabla_x D_w^*(x)|| = 0$, and hence we have $||y_T(x) - x|| = 0, \forall x \sim P_s(x)$. So, we have $P_s(x) = P_r(y)$. Therefore, $\theta$ also reaches its optimum $\theta^*$, and $(w, \theta)$ is an equilibrium point.

In practice, WGAN-QC is performing an alternating gradient descent between the discriminator and generator whereas the update operator in Eq. (15) is for simultaneous gradient descent. Therefore, we provide the following lemma to build the connection between the simultaneous and alternating gradient descent[1]

Lemma 4

In a local neighbourhoud of an equilibrium point, for alternating gradient descent of the discriminator updating $k_D > 0$ times and the generator updating $k_G$ times per alternating round of WGAN-QC with small enough learning rate $\alpha > 0$, all the eigenvalues of the Jacobian of the update operator $\hat{U} = \hat{U}_\theta^{k_G} \circ \hat{U}_w^{k_D}$ will be arbitrarily close to those of the Jacobian of the gradient operator $U(w, \theta)$ of simultaneous gradient descent at $(w^*, \theta^*)$, where $\hat{U}_w$ on $(w, \theta)$ is defined as

$$\hat{U}_w(w, \theta) = \left( w - \alpha \nabla_w \mathcal{L}_D(w, \theta) \right)$$

(20)

[1] This generalizes Lemma A.5 in [7]
and $\hat{U}_\theta$ on $(w, \theta)$ is defined as

$$\hat{U}_\theta (w, \theta) = (\theta - \alpha \nabla_\theta \mathcal{L}_G(w, \theta))$$

(21)

$\hat{U}_\theta^k (w, \theta)$ and $\hat{U}_\theta^k (w, \theta)$ mean performing $\hat{U}_\theta (w, \theta)$ and $\hat{U}_\theta (w, \theta)$, $k_D$ and $k_G$ times, respectively.

**Proof.** We adopt a similar proof to that of Lemma A.5 in [7]. The Jacobian of $U(w, \theta)$ is expressed as:

$$U(w, \theta)' = I - \alpha g'(w, \theta)$$

where $g'(w, \theta)$ is defined in Eq. (16) in the main paper. Therefore, $U(w^*, \theta^*)' = I - \alpha g'(w^*, \theta^*)$

$$\hat{U}(w^*, \theta^*)' = \hat{U}_\theta'(w^*, \theta^*)^{k_G} \cdot \hat{U}_\theta'(w^*, \theta^*)^{k_D} = (I - \alpha g'_\theta(w^*, \theta^*))^{k_G} \cdot (I - \alpha g'_\theta(w^*, \theta^*))^{k_D}$$

(23)

where

$$g'_\theta(w^*, \theta^*) = \begin{pmatrix} \nabla^2_w \mathcal{L}_D(w^*, \theta^*) & 0 \\ 0 & 0 \end{pmatrix}$$

(24)

and

$$g'_\theta(w^*, \theta^*) = \begin{pmatrix} 0 & 0 \\ 0 & \nabla^2_\theta \mathcal{L}_G(w^*, \theta^*) \end{pmatrix}$$

(25)

For small enough learning rate $\alpha$, we have

$$\hat{U}(w^*, \theta^*)' = I - \alpha (g'_\theta(w^*, \theta^*) + g'_\theta(w^*, \theta^*)) + O(\alpha^2)I$$

(26)

At the equilibrium point, $g'_\theta(w^*, \theta^*) = 0$ and $g'(w^*, \theta^*)$ in Lemma 2 has the same eigenvalues as $g'_\theta(w^*, \theta^*)$. Hence, for small enough learning rates, all the eigenvalues of $\hat{U} = \hat{U}_\theta^{k_G} \circ \hat{U}_\theta^{k_D}$ will be arbitrarily close to those of $U(w, \theta)$ at the equilibrium point $(w^*, \theta^*)$.

\[ \square \]

**Theorem 2** in Section 4

Suppose Assumptions 1 and 2 are satisfied, then for small enough learning rate $\alpha$, there exists $\lambda$ such that WGAN-QC converges to a local equilibrium point.

**Proof.** According to Lemma 4, for small enough learning rate, the eigenvalues of the Jacobian of the alternating gradient update operator are arbitrarily close to that of the simultaneous gradient update operator. Hence we analyze the Jacobian of simultaneous gradient of WGAN-QC.

At an equilibrium point, all the eigenvalues of the Jacobian of the gradient field $g(w, \theta)$ are non-negative, and the positive ones are the same as the eigenvalues of the Hessian, $M_{DD} + M_R$, of the discriminator loss. This allows us to study the convergence of only $w$ rather than $(w, \theta)$ near equilibrium. Since $M_{DD} + M_R$ is positive definite by Lemma 2, for simultaneous gradient descent, the discriminator parameters $w$ converge to optimum.

According to Lemma 3, when $w$ converges to the optimum, $\theta$ also converges to its optimum. Therefore, $(w, \theta)$ converge to a local equilibrium point.

\[ \square \]

3. Experiments

3.1. Experimental Settings

In all the experiments, we preform one discriminator iteration per generator iteration for WGAN-QC. On the Dirac distribution experiment, we set $\gamma = 10$ in CRGAN, $k = 2$ and $p = 6$ in WGAN-div as suggested in the original papers. We use gradient descent for all the methods and the learning rate for all the methods is 0.01.

On the CelebA and the LSUN datasets, we crop the image size to 64x64. On the CelebA-HQ dataset, we resize the image size to 256x256. On the ImageNet dog dataset, we crop the image size to 128x128. For WGAN-GP [2], WGAN-div [10] and CRGAN [7], we use the default parameters and architectures suggested in their papers. As suggested in the WGAN-div paper [10], we perform 100,000 and 200,000 iterations for WGAN-div on the CelebA and LSUN datasets, respectively. We perform the same numbers of iterations as WGAN-div on the two datasets for WGAN-GP and CRGAN. For WGAN-QC, we perform 60,000, 125,000, 100,000 and 150,000 iterations on the CelebA, CelebA-HQ, LSUN and ImageNet dog datasets,
respectively. On these real world datasets, we use the Adam optimizer [5] for WGAN-QC in all the experiments. Learning rate, weights $\beta_1$ and $\beta_2$ in Adam on all the datasets are set to $1e^{-4}$, 0.5 and 0.999, respectively. We decay the learning rate to $1e^{-5}$ starting from the 120,000 iteration on the ImageNet dog dataset for WGAN-QC. Weight $\gamma$ for WGAN-QC is set to 0.1 on the CelebA [6] and LSUN [11] datasets, and set to 0.05 on the CelebA-HQ dataset [4]. We use the exponential moving average method [12] on the generator with a decay of 0.999 during the last 5000 iterations for WGAN-QC.

### 3.2. Network Architectures

The architectures of the discriminators and generators in WGAN-QC for image sizes $64 \times 64$, $128 \times 128$ and $256 \times 256$ are shown in Tables 1–6. Same as the CRGAN paper [7], we multiply the residual part by 0.1 in every ResNet Block. We use leaky ReLU as the non-linear activation. We use batch normalization [3] only for the generator. The batch size for the CelebA and LSUN datasets is set to 64, for the CelebA-HQ dataset is set to 16 because of memory constraints.

#### Table 1. The discriminator architecture for WGAN-QC $64 \times 64$ image size.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Filter size</th>
<th>Resampling</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td>3 × 3</td>
<td>-</td>
<td>64 × 64 × 64</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>128 × 32 × 32</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>256 × 16 × 16</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>512 × 8 × 8</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Table 2. The generator architecture for WGAN-QC $64 \times 64$ image size.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Filter size</th>
<th>Resampling</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>-</td>
<td>-</td>
<td>128</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Up</td>
<td>512 × 8 × 8</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Up</td>
<td>256 × 16 × 16</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Up</td>
<td>128 × 32 × 32</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Up</td>
<td>64 × 64 × 64</td>
</tr>
<tr>
<td>Conv, tanh</td>
<td>-</td>
<td>-</td>
<td>3 × 64 × 64</td>
</tr>
</tbody>
</table>

#### Table 3. The discriminator architecture for WGAN-QC $128 \times 128$ image size.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Filter size</th>
<th>Resampling</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td>3 × 3</td>
<td>-</td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>256 × 32 × 32</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>512 × 16 × 16</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>512 × 8 × 8</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>Down</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3 × 3] × 2</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4. The generator architecture for WGAN-QC 128×128 image size.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Filter size</th>
<th>Resampling</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>-</td>
<td>-</td>
<td>128</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3×3] × 2</td>
<td>Up</td>
<td>512 × 8 × 8</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3×3] × 2</td>
<td>Up</td>
<td>512 × 16 × 16</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3×3] × 2</td>
<td>Up</td>
<td>256 × 32 × 32</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3×3] × 2</td>
<td>Up</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>[3×3] × 2</td>
<td>Up</td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>Conv, tanh</td>
<td>-</td>
<td>-</td>
<td>3 × 128 × 128</td>
</tr>
</tbody>
</table>

Table 5. The discriminator architecture for WGAN-QC 256×256 image size.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Filter size</th>
<th>Resampling</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv</td>
<td>3×3</td>
<td>-</td>
<td>64 × 256 × 256</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Down</td>
<td>128 × 128 × 128</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Down</td>
<td>256 × 64 × 64</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Down</td>
<td>512 × 32 × 32</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Down</td>
<td>512 × 16 × 16</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Down</td>
<td>512 × 8 × 8</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Down</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6. The generator architecture for WGAN-QC 256×256 image size.

<table>
<thead>
<tr>
<th>Discriminator</th>
<th>Filter size</th>
<th>Resampling</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>-</td>
<td>-</td>
<td>128</td>
</tr>
<tr>
<td>Linear</td>
<td>-</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>-</td>
<td>512 × 4 × 4</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Up</td>
<td>512 × 8 × 8</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Up</td>
<td>512 × 16 × 16</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Up</td>
<td>512 × 32 × 32</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Up</td>
<td>256 × 64 × 64</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Up</td>
<td>128 × 128 × 128</td>
</tr>
<tr>
<td>ResNet Block</td>
<td>3×3</td>
<td>Up</td>
<td>64 × 256 × 256</td>
</tr>
<tr>
<td>Conv, tanh</td>
<td>-</td>
<td>-</td>
<td>3 × 256 × 256</td>
</tr>
</tbody>
</table>

3.3. Results

We show larger versions of the figures in the main paper here. We show more randomly generated images and interpolation results on the CelebA-HQ dataset by WGAN-QC (Figs. 5 and 6). All images shown in Figs. 1-4 and Figs. 7-10 are 64×64 images. All images shown in Figs. 5 and 6 are 256×256 images.
Figure 1. Larger version of Figure 3 (a) in the main paper (WGAN-GP). Obvious failure cases are marked with red boxes.
Figure 2. Larger version of Figure 3 (b) in the main paper (WGAN-div). Obvious failure cases are marked with red boxes.
Figure 3. Larger version of Figure 3 (c) in the main paper (CRGAN). Red boxes in (c) suggest the mode collapse problem of CRGAN.
Figure 4. Larger version of Figure 3 (d) in the main paper (WGAN-QC). All images generated by WGAN-QC are complete, smooth and distinct from each other.
Figure 5. Face images randomly generated by WGAN-QC on the CelebA HQ dataset (image size is 256×256).
Figure 6. Face interpolation by WGAN-QC on the CelebA-HQ dataset. Transitions between the leftmost and rightmost faces appear smooth and plausible.
Figure 7. Larger version of Figure 6 (a) in the main paper (WGAN-GP). Obvious failure cases are marked with red boxes.
Figure 8. Larger version of Figure 6 (b) in the main paper (WGAN-div). Obvious failure cases are marked with red boxes.
Figure 9. Larger version of Figure 6 (c) in the main paper (CRGAN). Obvious failure cases are marked with red boxes.
Figure 10. Larger version of Figure 6 (d) in the main paper (WGAN-QC). Obvious failure cases are marked with red boxes.
References


