Wasserstein GAN with Quadratic Transport Cost
Supplementary Material

Huidong Liu, Xianfeng Gu, Dimitris Samaras
Stony Brook University
Stony Brook, NY 11794, USA
{huidliu, gu, samaras}@cs.stonybrook.edu

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:

$$\begin{aligned} \max_{H_i, H_j} \quad & \frac{1}{m}\sum_{i\in\mathcal{I}} H_i - \frac{1}{n}\sum_{j\in\mathcal{J}} H_j \\ \text{s.t.} \quad & H_i - H_j \le c(x_j, y_i), \\ & \forall i \in \mathcal{I},\ \forall j \in \mathcal{J} \end{aligned} \tag{1}$$

where $\mathcal{I}$ and $\mathcal{J}$ are disjoint sets, then for each x_j, there exists a $t \in \mathcal{I}$, such that $H^*_t - H^*_j = c(x_j, y_t)$.

We prove this by contradiction, i.e., there exists one x_s, $s \in \mathcal{J}$, such that we cannot find a y_i such that $H^*_i - H^*_s = c(x_s, y_i)$, $\forall i \in \mathcal{I}$. This means that $H^*_s > \sup_{i\in\mathcal{I}}\{H^*_i - c(x_s, y_i)\}$. We can construct H', such that $H'_i = H^*_i, \forall i \in \mathcal{I}$, $H'_j = H^*_j, \forall j \in \mathcal{J} - \{s\}$, and $H'_s = \sup_{i\in\mathcal{I}}\{H^*_i - c(x_s, y_i)\}$. 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, $t \in \mathcal{I}$, such that $H^*_t - H^*_j = c(x_j, y_t)$.

Next, we will show that if $H^*_t - H^*_j = c(x_j, y_t)$, then $t = \sigma(j)$. Since $c(x_j, y_t) = \frac{K}{2}||x_j - y_t||_2^2$, we have $\frac{K}{2}||x_j - y_t||_2^2 + H^*_j - H^*_i = 0$. According to the definition of $\sigma(j)$ in Eq. (6), and the fact that $H^*_t - H^*_j \le c(x_j, y_t)$, we have $t = \sigma(j)$. □

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 [1], can also be used to prove Eq. (8). First we will introduce the following theorem.

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

$$T(x) = x + (\nabla h)^{-1}(\nabla D(x)) \tag{2}$$

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 σ in Eq. (6), we know that x_j is transported to $y_{\sigma(j)}$ [9], i.e., $T(x_j) = y_{\sigma(j)}$. If $c(x, y) = \frac{K}{2}||x - y||_2^2$, then

$$T(x) = x + \frac{1}{K}\nabla D(x) \tag{3}$$

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),

$$\frac{1}{m}\sum_{i\in\mathcal{I}} D_w^*(y_i) - \frac{1}{n}\sum_{j\in\mathcal{J}} D_w^*(x_j) \tag{6}$$

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., $\frac{1}{m}\sum_{i\in\mathcal{I}} D_w^*(y_i) = \frac{1}{m}\sum_{i\in\mathcal{I}} H_i^*$, and $D_w^*(x_j) = H_j^*, \forall j \in \mathcal{J}$. Since $\frac{1}{m}\sum_{i\in\mathcal{I}} H_i^* - \frac{1}{n}\sum_{j\in\mathcal{J}} H_j^*$ is the quadratic Wasserstein distance between $\hat{X}$ and $\hat{Y}$, $\frac{1}{m}\sum_{i\in\mathcal{I}} D_w^*(y_i) - \frac{1}{n}\sum_{j\in\mathcal{J}} D_w^*(x_j)$ is the quadratic Wasserstein distance between $\hat{X}$ and $\hat{Y}$. □

Lemma 1 in Section 4

The second order derivative of the regularization term

$$\mathcal{L}_R = \frac{\lambda}{2}\mathbb{E}_{\mathbb{P}_s(x)}[\big(||\nabla_x D_w(x)|| - ||y_{T(x)} - x||\big)^2] \tag{7}$$

with respect to (w, θ) at the equilibrium point is given by:

$$M_R = \lambda \cdot \mathbb{E}_{\mathbb{P}_s(x)}[\nabla_{w,x} D_{w^*}(x) \nabla_{w,x} D_{w^*}(x)^\intercal] \tag{8}$$

Proof. At the equilibrium point (w^*, θ^*), every point x is transported to $y_{T(x)}$, and $x = y_{T(x)}$. Therefore, at the equilibrium point, the regularization term is

$$\mathcal{L}_R' = \frac{\lambda}{2}\mathbb{E}_{\mathbb{P}_s(x)}[||\nabla_x D_w(x)||^2] \tag{9}$$

The first order derivative at the equilibrium point is $\nabla_w \mathcal{L}_R' = \lambda \cdot \mathbb{E}_{\mathbb{P}_s(x)}[\nabla_{w,x} D_{w^*}(x) \nabla_x D_{w^*}(x)]$. 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 $\mathbb{P}_s(x) = \mathbb{P}_r(y)$ at the equilibrium point), we have $M_R = \lambda \cdot \mathbb{E}_{\mathbb{P}_s(x)}[\nabla_{w,x} D_{w^*}(x) \nabla_{w,x} D_{w^*}(x)^\intercal]$ □

Lemma 2 in Section 4

The Jacobian of the gradient field g(w, θ) at the equilibrium point (w^*, θ^*) 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,

$$M_{DD} = \mathbb{E}_{\mathbb{P}_r(y)}[\nabla_w D_{w^*}(y)] \cdot \mathbb{E}_{\mathbb{P}_r(y)}[\nabla_w D_{w^*}(y)^\intercal] + \mathbb{E}_{\mathbb{P}_r(y)}[\nabla_w D_{w^*}(y) \nabla_w D_{w^*}(y)^\intercal] \tag{11}$$

,

$$M_{GD} = -\mathbb{E}_{\mathbb{P}_s(x)}[\nabla^2_{w,x} D_{w^*}(x) \nabla_\theta G_{\theta^*}(z)^\intercal] \tag{12}$$

and $M_{DD} + M_R$ is positive definite.

Proof. We give the first order derivatives of $\mathcal{L}_D$ and $\mathcal{L}_G$ below:

$$\nabla_w \mathcal{L}_D(w,\theta) = \big(\mathbb{E}_{\mathbb{P}_r(y)}[D_w(y)] - \mathbb{E}_{\mathbb{P}_r(y)}[H_r(y)]\big) \cdot \mathbb{E}_{\mathbb{P}_r(y)}[\nabla_w D_w(y)] + \mathbb{E}_{\mathbb{P}_s(x)}[(D_w(x) - H_s(x))\nabla_w D_w(x)] + \nabla_w \mathcal{L}_R \tag{13}$$

$$\nabla_\theta \mathcal{L}_G(w,\theta) = -\mathbb{E}_{\mathbb{P}(z)}[\nabla_\theta G_\theta(z) \nabla_x D_w(x)] \tag{14}$$

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

$$\begin{aligned} &\nabla^2_w \mathcal{L}_D(w,\theta) \\ &= \big(\mathbb{E}_{\mathbb{P}_r(y)}[D_w(y)] - \mathbb{E}_{\mathbb{P}_r(y)}[H_r(y)]\big) \cdot \mathbb{E}_{\mathbb{P}_r(y)}[\nabla^2_w D_w(y)] + \mathbb{E}_{\mathbb{P}_r(y)}[\nabla_w D_w(y)] \cdot \mathbb{E}_{\mathbb{P}_r(y)}[\nabla_w D_w(y)^\intercal] \\ &+ \mathbb{E}_{\mathbb{P}_s(x)}[(D_w(x) - H_s(x))\nabla^2_w D_w(x)] + \mathbb{E}_{\mathbb{P}_s(x)}[\nabla_w D_w(x) \nabla_w D_w(x)^\intercal] + \nabla^2_w \mathcal{L}_R \end{aligned} \tag{15}$$

$$\nabla^2_{\theta,w} \mathcal{L}_D(w,\theta) = \mathbb{E}_{\mathbb{P}_s(x)}[(D_w(x) - H_s(x))\nabla_\theta G_\theta(z) \nabla^2_{x,w} D_w(x)] + \mathbb{E}_{\mathbb{P}_s(x)}[\nabla_\theta G(z) \nabla_x D_w(x) \nabla_w D_w(x)^\intercal] \tag{16}$$

$$\nabla^2_{w,\theta} \mathcal{L}_G(w,\theta) = -\mathbb{E}_{\mathbb{P}_s(x)}[\nabla^2_{w,x} D_w(x) \nabla_\theta G_\theta(z)^\intercal] \tag{17}$$

$$\nabla^2_\theta \mathcal{L}_G(w,\theta) = -\mathbb{E}_{\mathbb{P}(z)}[\nabla_\theta G_\theta(z) \nabla^2_x D_w(x) \nabla_\theta G_\theta(z)^\intercal] - \mathbb{E}_{\mathbb{P}(z)}[\nabla^2_\theta G_\theta(z) \nabla_x D_w(x)] \tag{18}$$

Since at equilibrium, $D_{w^*}(x) = D_{w^*}(y) = 0$, $\nabla_x D_{w^*}(x) = 0$, $D_{w^*}(x) = H_s(x)$, $D_{w^*}(y) = H_r(y)$, $x = y_{T(x)}$ and $\nabla^2_w \mathcal{L}_R = M_R$, we have $\nabla^2_w \mathcal{L}_D(w^*,\theta^*) = M_{DD} + M_R$, $\nabla^2_{\theta,w} \mathcal{L}_D(w^*,\theta^*) = 0$, $\nabla^2_{w,\theta} \mathcal{L}_G(w^*,\theta^*) = M_{GD}$, and $\nabla^2_\theta \mathcal{L}_G(w^*,\theta^*) = 0$.

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, θ) in WGAN-QC using Eq. (15), if $w = w^*$, then $\theta = \theta^*$.

Proof. First we show that $||\nabla_x D_{w^*}(x)|| = 0, \forall x \sim \mathbb{P}_s(x)$. According to the definition of w^* in Eq. (17), we can find a θ^*, such that $\mathcal{L}_D(w^*,\theta^*) = 0$. According to the definition of θ^* in Eq. (17), we know that $||y_{T(x)} - x|| = 0, \forall x \sim \mathbb{P}_s(x)$. Therefore, $\mathbb{E}_{\mathbb{P}_s(x)}[||\nabla_x D_{w^*}(x)||_2^2] = 0$, and hence $||\nabla_x D_{w^*}(x)|| = 0, \forall x \sim \mathbb{P}_s(x)$.

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 θ. Thus, $\mathcal{L}_D(w^*,\theta) = 0$, and we have

$$\mathbb{E}_{\mathbb{P}_s(x)}[\big(||\nabla_x D_{w^*}(x)|| - K||y_{T(x)} - x||\big)^2] = 0 \tag{19}$$

Recall that $||\nabla_x D_{w^*}(x)|| = 0$, and hence we have $||y_{T(x)} - x|| = 0, \forall x \sim \mathbb{P}_s(x)$. So, we have $\mathbb{P}_s(x) = \mathbb{P}_r(y)$. Therefore, θ also reaches its optimum θ^*, and (w, θ) 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 neighbourhood 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^*,θ^*), where $\hat{U}_w$ on (w,θ) is defined as

$$\hat{U}_w(w,\theta) = \begin{pmatrix} w - \alpha \nabla_w \mathcal{L}_D(w,\theta) \\ \theta \end{pmatrix} \tag{20}$$

[1] This generalizes Lemma A.5 in [7]

and $\hat{U}_\theta$ on (w, θ) is defined as

$$\hat{U}_\theta(w,\theta) = \begin{pmatrix} w \\ \theta - \alpha\nabla_\theta\mathcal{L}_G(w,\theta) \end{pmatrix} \tag{21}$$

$\hat{U}_\theta^{k_D}(w,\theta)$ and $\hat{U}_\theta^{k_G}(w,\theta)$ 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:

$$U(w,\theta)' = I - \alpha g'(w,\theta) \tag{22}$$

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

$$\hat{U}(w^*,\theta^*)' = \quad \hat{U}'_\theta(w^*,\theta^*)^{k_G} \cdot \hat{U}'_w(w^*,\theta^*)^{k_D} = \quad (I - \alpha g'_\theta(w^*,\theta^*))^{k_G} \cdot (I - \alpha g'_w(w^*,\theta^*))^{k_D} \tag{23}$$

where

$$g'_w(w^*,\theta^*) = \begin{pmatrix} \nabla_w^2\mathcal{L}_D(w^*,\theta^*) & 0 \\ 0 & 0 \end{pmatrix} \tag{24}$$

and

$$g'_\theta(w^*,\theta^*) = \begin{pmatrix} 0 & 0 \\ 0 & \nabla_\theta^2\mathcal{L}_G(w^*,\theta^*) \end{pmatrix} \tag{25}$$

For small enough learning rate α, we have

$$\hat{U}(w^*,\theta^*)' = I - \alpha(g'_\theta(w^*,\theta^*) + g'_w(w^*,\theta^*)) + \mathcal{O}(\alpha^2)I \tag{26}$$

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

□

Theorem 2 in Section 4

Suppose Assumptions 1 and 2 are satisfied, then for small enough learning rate α, there exists λ 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,θ) 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, θ also converges to its optimum. Therefore, (w, θ) converge to a local equilibrium point. □

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 datsets, we crop the image size to 64×64. On the CelebA-HQ dataset, we resize the image size to 256×256. On the ImageNet dog dataset, we crop the image size to 128×128. 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 β_1 and β_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 γ 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 × 64, 128 × 128 and 256 × 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 64×64 image size.

Discriminator	Filter size	Resampling	Output size
Conv	3×3	-	64 × 64 × 64
ResNet Block	[3×3] × 2	Down	128 × 32 × 32
ResNet Block	[3×3] × 2	Down	256 × 16 × 16
ResNet Block	[3×3] × 2	Down	512 × 8 × 8
ResNet Block	[3×3] × 2	Down	512 × 4 × 4
ResNet Block	[3×3] × 2	-	512 × 4 × 4
Linear	-	-	1

Table 2. The generator architecture for WGAN-QC 64×64 image size.

Discriminator	Filter size	Resampling	Output size
Noise	-	-	128
Linear	-	-	512 × 4 × 4
ResNet Block	[3×3] × 2	-	512 × 4 × 4
ResNet Block	[3×3] × 2	Up	512 × 8 × 8
ResNet Block	[3×3] × 2	Up	256 × 16 × 16
ResNet Block	[3×3] × 2	Up	128 × 32 × 32
ResNet Block	[3×3] × 2	Up	64 × 64 × 64
Conv, tanh	-	-	3 × 64 × 64

Table 3. The discriminator architecture for WGAN-QC 128×128 image size.

Discriminator	Filter size	Resampling	Output size
Conv	3×3	-	64 × 128 × 128
ResNet Block	[3×3] × 2	Down	128 × 64 × 64
ResNet Block	[3×3] × 2	Down	256 × 32 × 32
ResNet Block	[3×3] × 2	Down	512 × 16 × 16
ResNet Block	[3×3] × 2	Down	512 × 8 × 8
ResNet Block	[3×3] × 2	Down	512 × 4 × 4
ResNet Block	[3×3] × 2	-	512 × 4 × 4
Linear	-	-	1

Table 4. The generator architecture for WGAN-QC 128×128 image size.

Discriminator	Filter size	Resampling	Output size
Noise	-	-	128
Linear	-	-	512 × 4 × 4
ResNet Block	[3×3] × 2	-	512 × 4 × 4
ResNet Block	[3×3] × 2	Up	512 × 8 × 8
ResNet Block	[3×3] × 2	Up	512 × 16 × 16
ResNet Block	[3×3] × 2	Up	256 × 32 × 32
ResNet Block	[3×3] × 2	Up	128 × 64 × 64
ResNet Block	[3×3] × 2	Up	64 × 128 × 128
Conv, tanh	-	-	3 × 128 × 128

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

Discriminator	Filter size	Resampling	Output size
Conv	3×3	-	64 × 256 × 256
ResNet Block	3×3	Down	128 × 128 × 128
ResNet Block	3×3	Down	256 × 64 × 64
ResNet Block	3×3	Down	512 × 32 × 32
ResNet Block	3×3	Down	512 × 16 × 16
ResNet Block	3×3	Down	512 × 8 × 8
ResNet Block	3×3	Down	512 × 4 × 4
ResNet Block	3×3	-	512 × 4 × 4
Linear	-	-	1

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

Discriminator	Filter size	Resampling	Output size
Noise	-	-	128
Linear	-	-	512 × 4 × 4
ResNet Block	3×3	-	512 × 4 × 4
ResNet Block	3×3	Up	512 × 8 × 8
ResNet Block	3×3	Up	512 × 16 × 16
ResNet Block	3×3	Up	512 × 32 × 32
ResNet Block	3×3	Up	256 × 64 × 64
ResNet Block	3×3	Up	128 × 128 × 128
ResNet Block	3×3	Up	64 × 256 × 256
Conv, tanh	-	-	3 × 256 × 256

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.

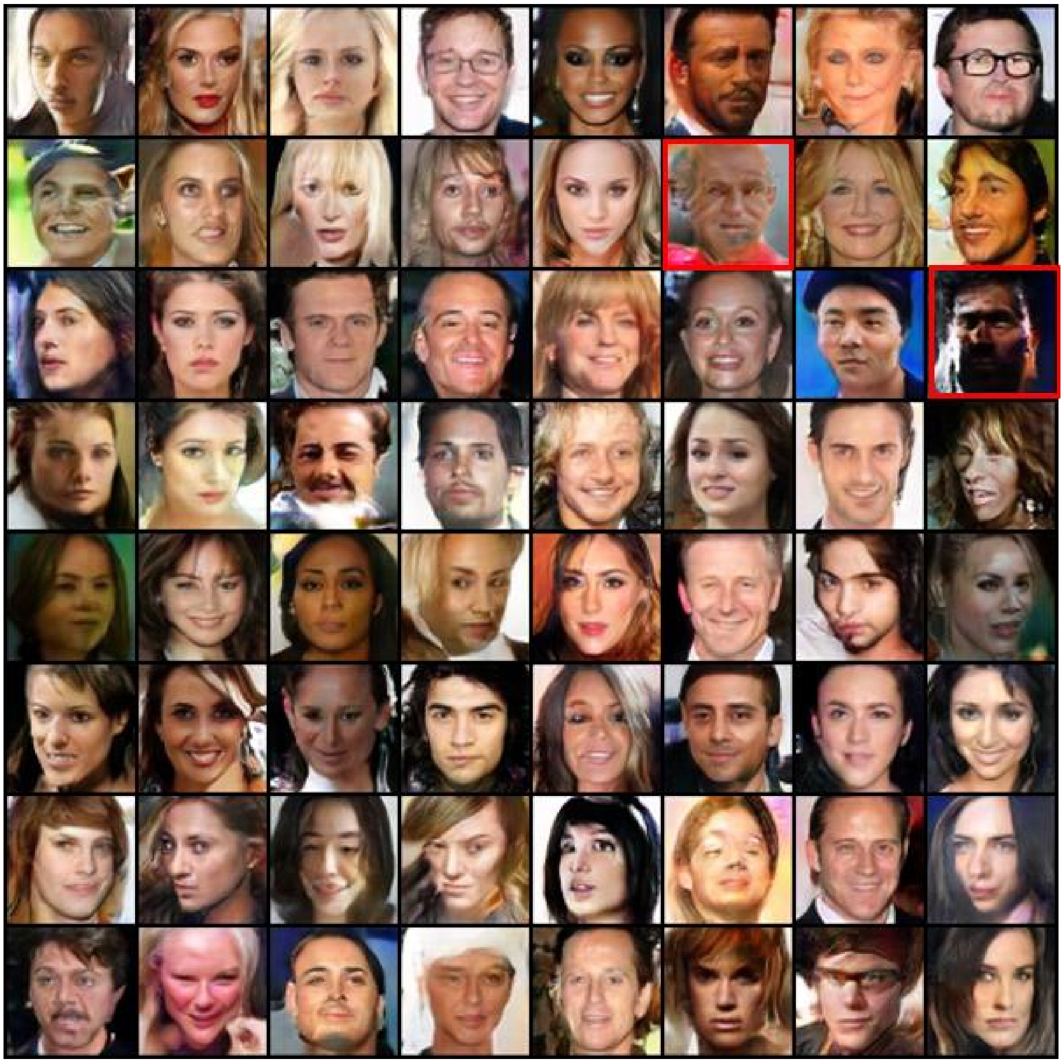

Figure 1. Larger version of Figure 3 (a) in the main paper (WGAN-GP). Obvious failure cases are marked with red boxes.

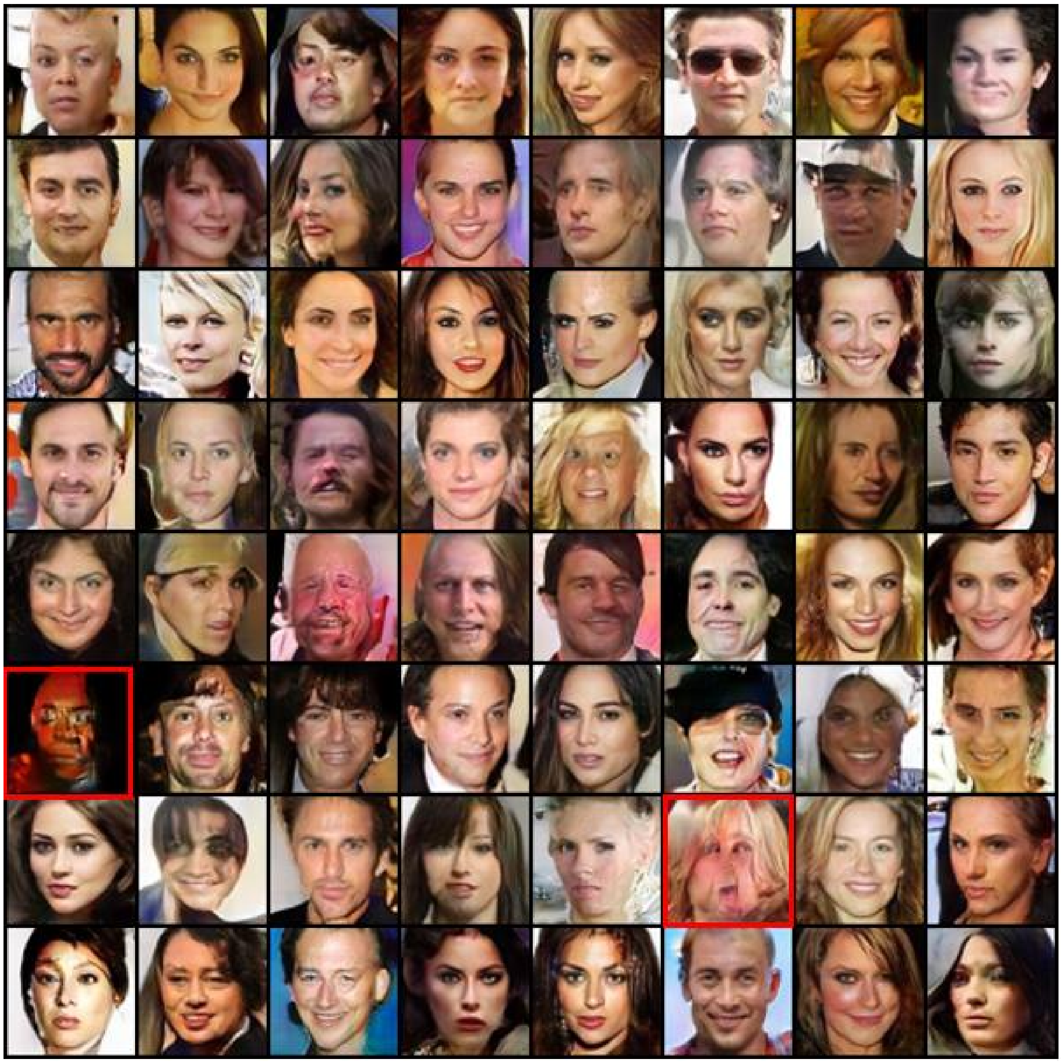

Figure 2. Larger version of Figure 3 (b) in the main paper (WGAN-div). Obvious failure cases are marked with red boxes.

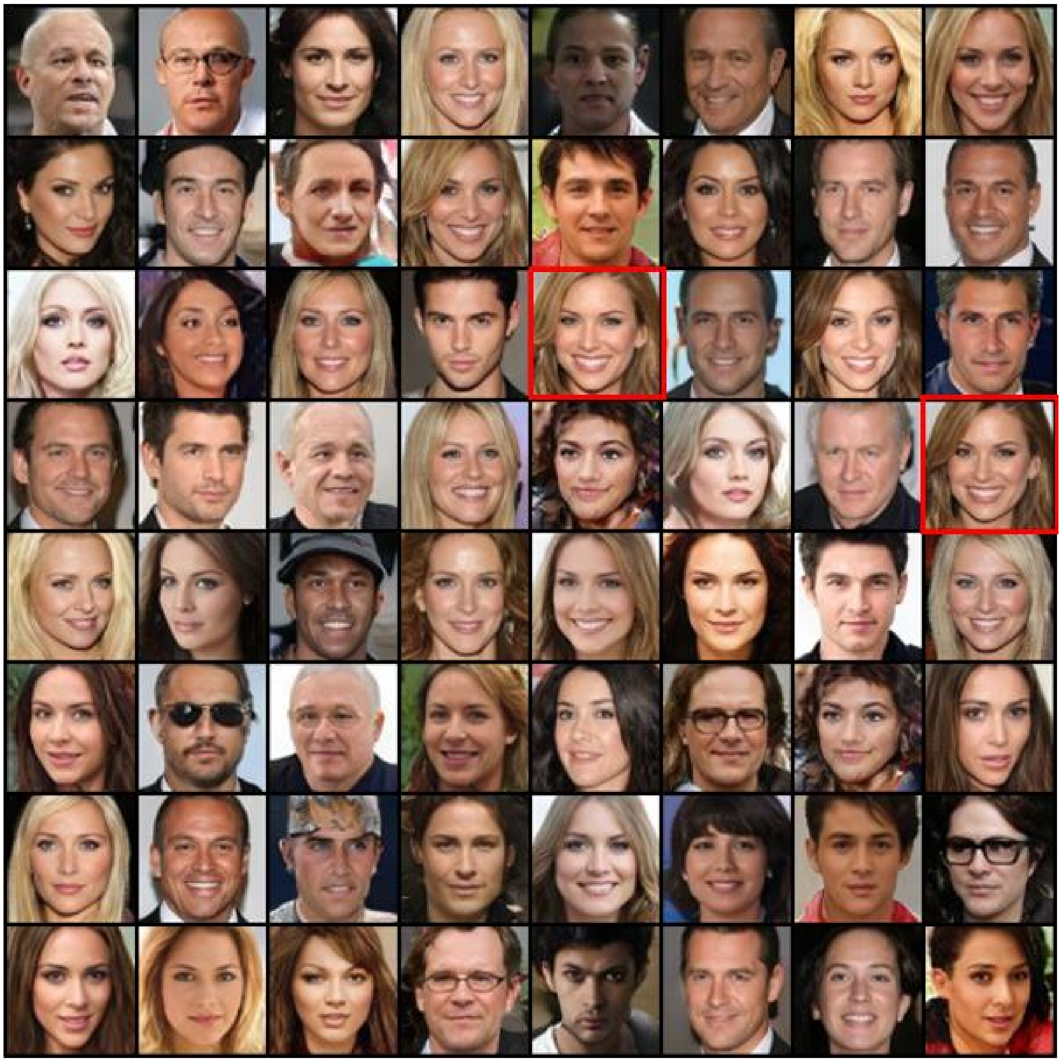

Figure 3. Larger version of Figure 3 (c) in the main paper (CRGAN). Red boxes in (c) suggest the mode collapse problem of CRGAN.

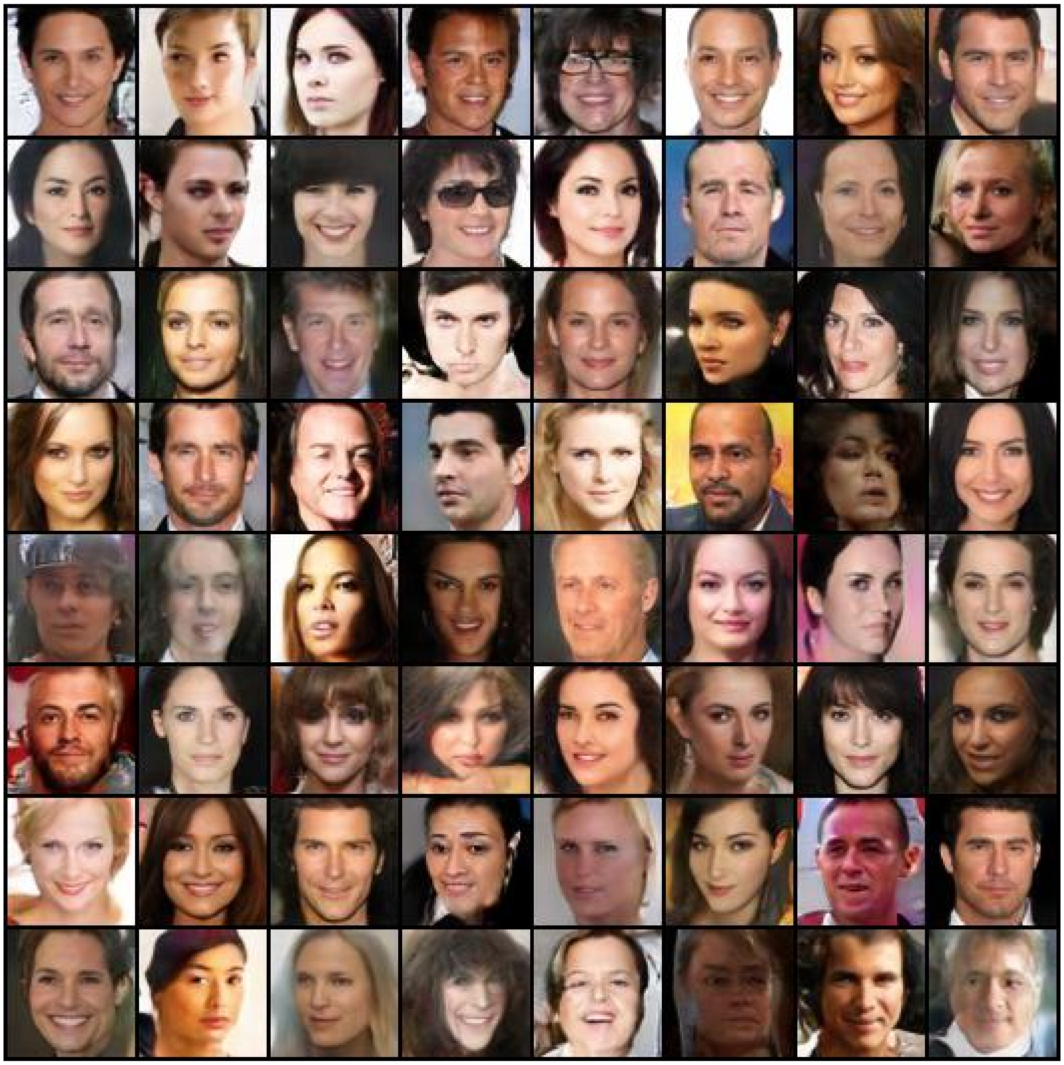

Figure 4. Larger version of Figure 3 (d) in the main paper (WGAN-QC). All images generated by WGAN-QC are complete, smooth and distinct from each other.

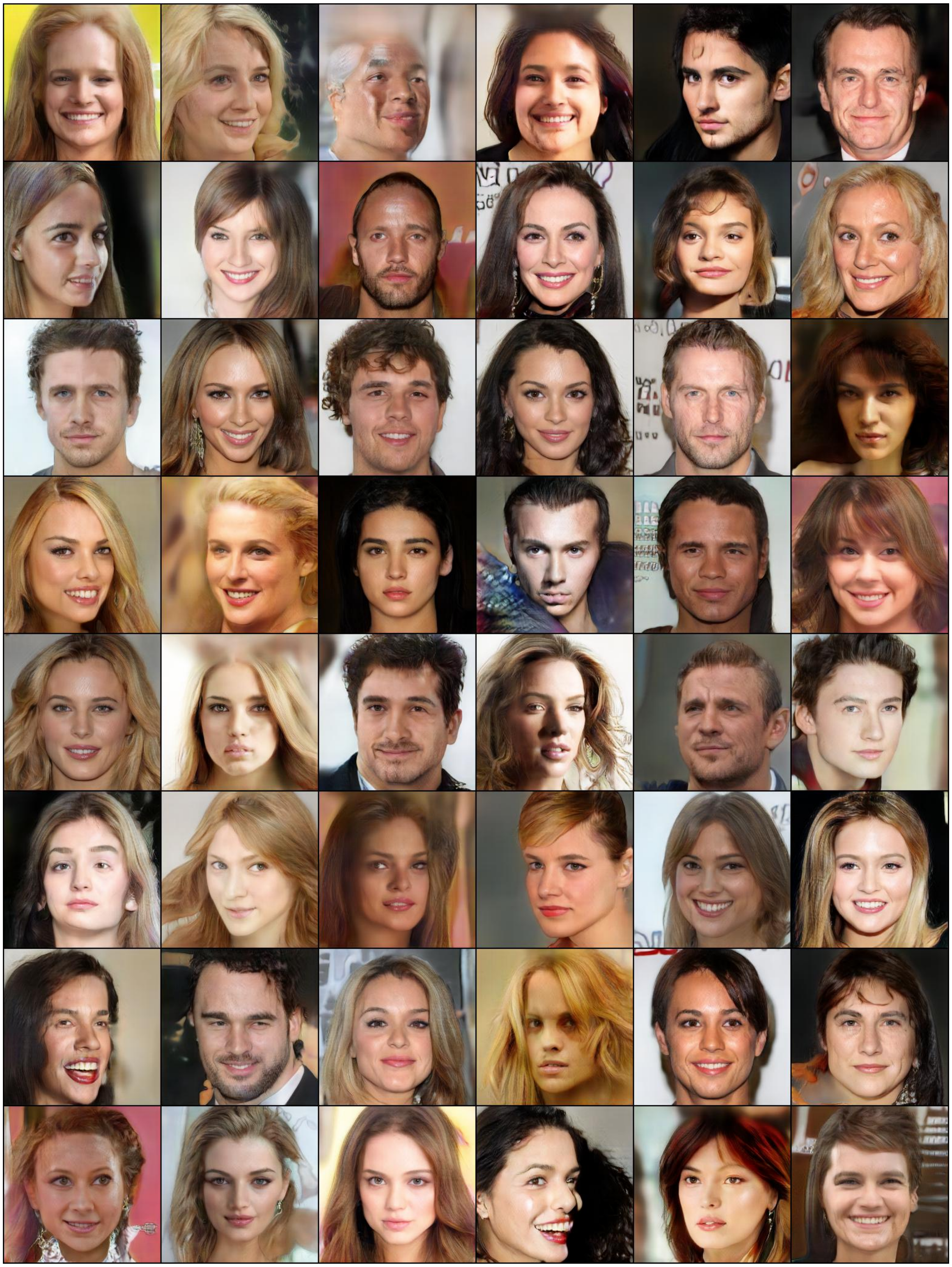

Figure 5. Face images randomly generated by WGAN-QC on the CelebA HQ dataset (image size is 256×256).

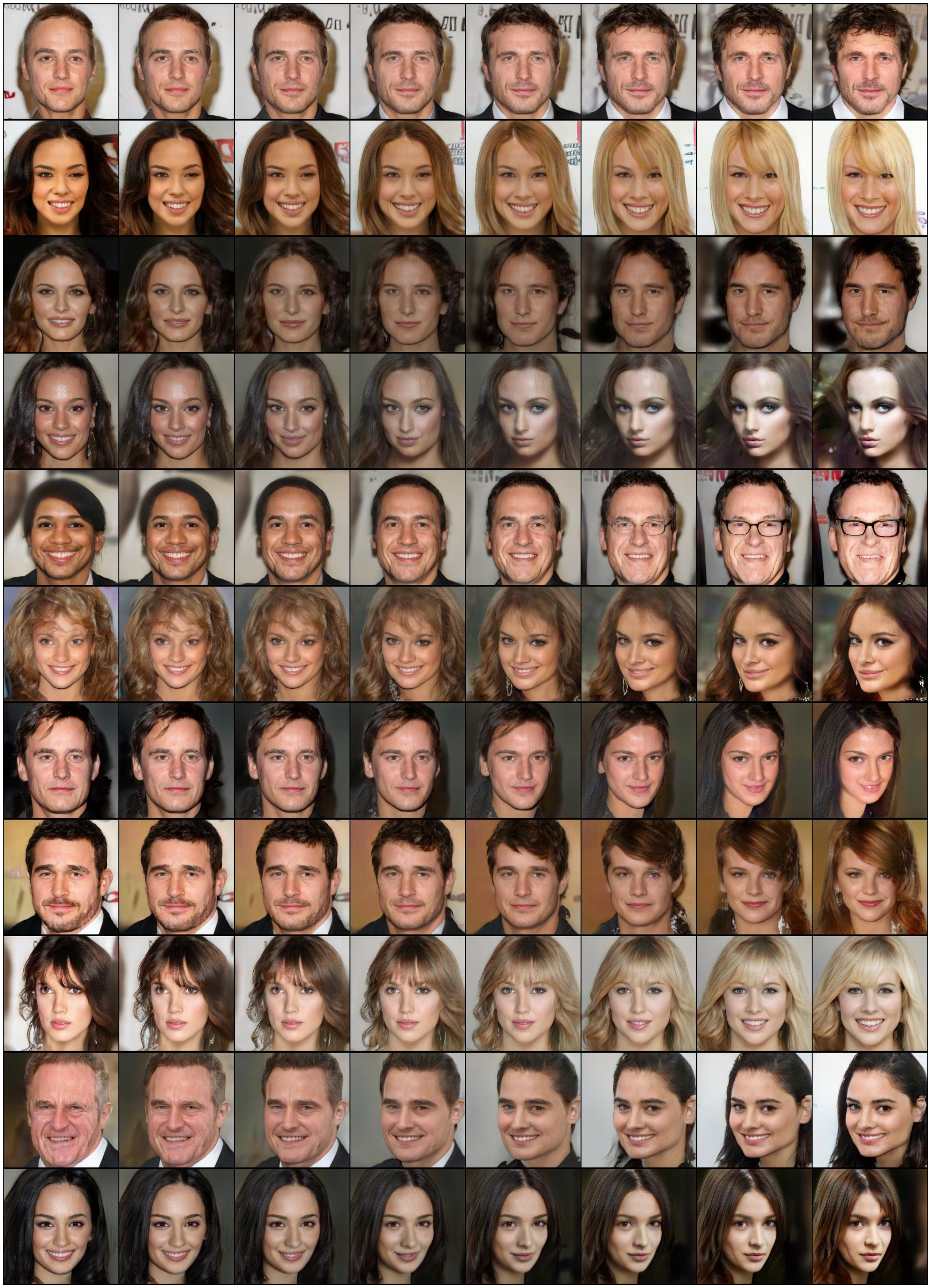

Figure 6. Face interpolation by WGAN-QC on the CelebA-HQ dataset. Transitions between the leftmost and rightmost faces appear smooth and plausible.

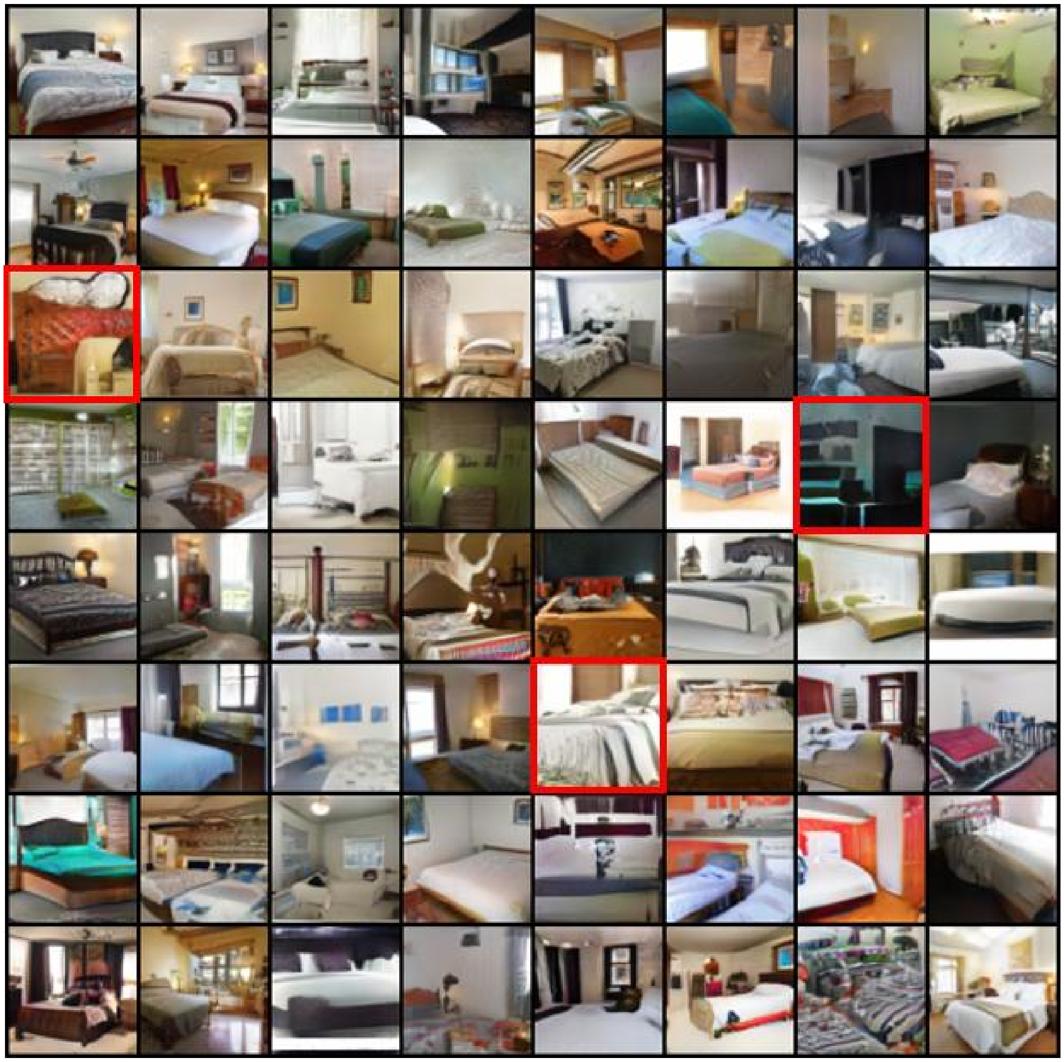

Figure 7. Larger version of Figure 6 (a) in the main paper (WGAN-GP). Obvious failure cases are marked with red boxes.

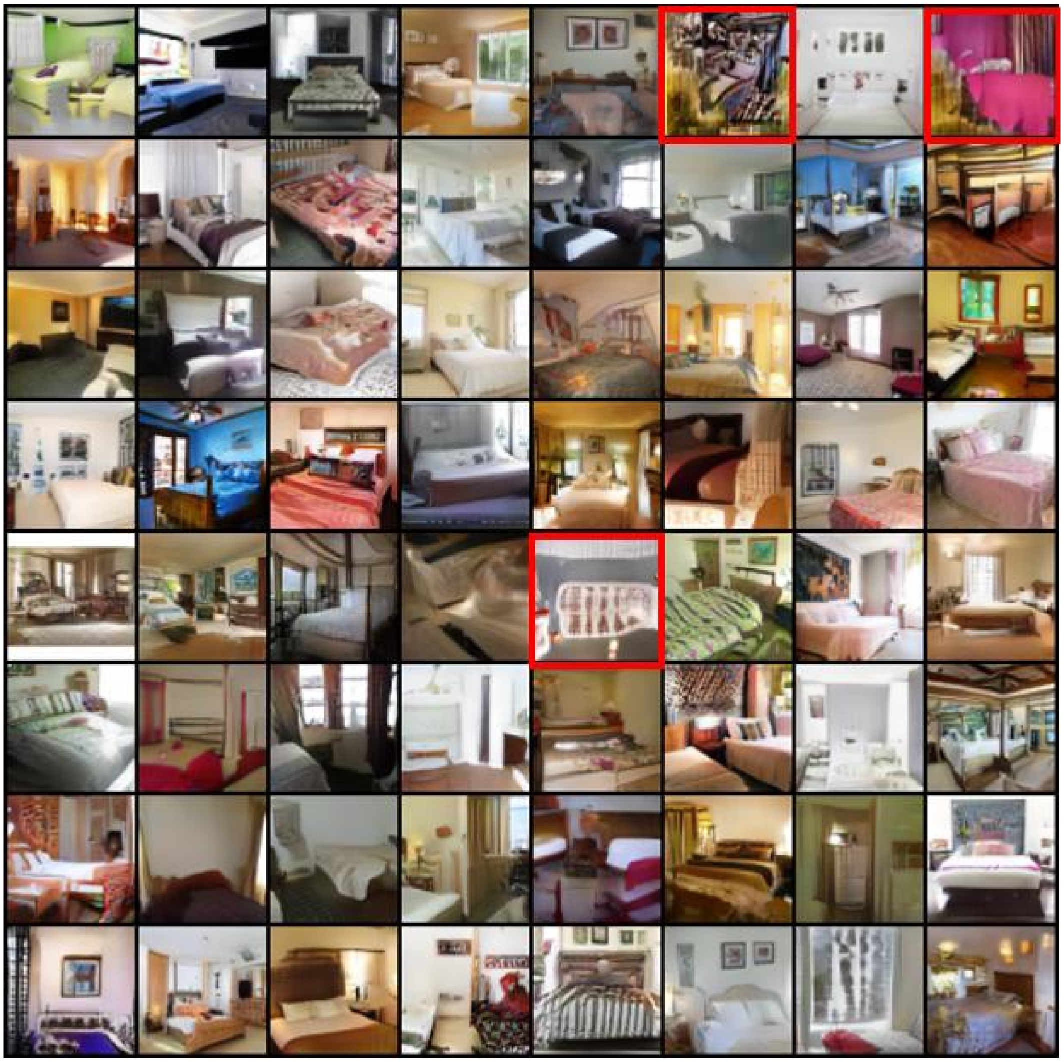

Figure 8. Larger version of Figure 6 (b) in the main paper (WGAN-div). Obvious failure cases are marked with red boxes.

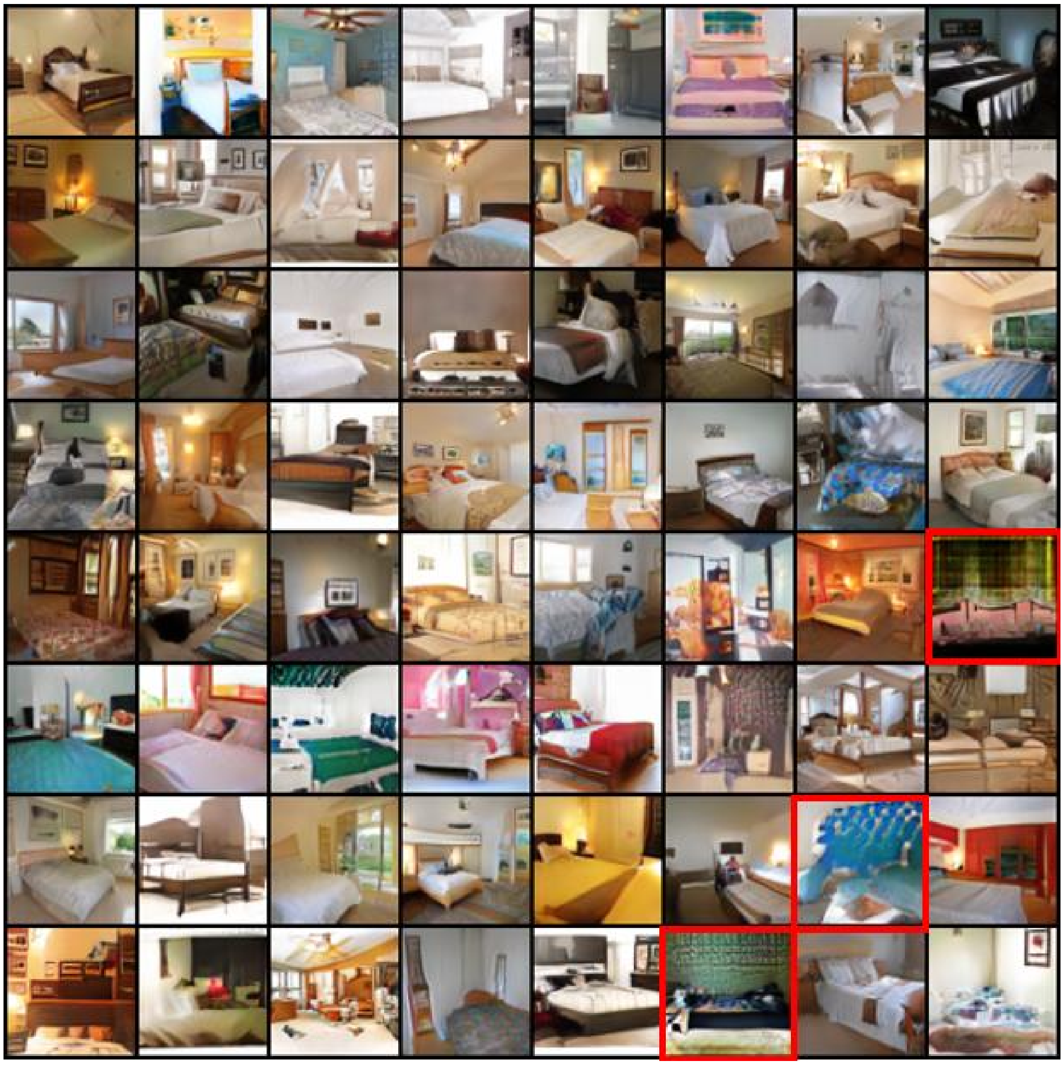

Figure 9. Larger version of Figure 6 (c) in the main paper (CRGAN). Obvious failure cases are marked with red boxes.

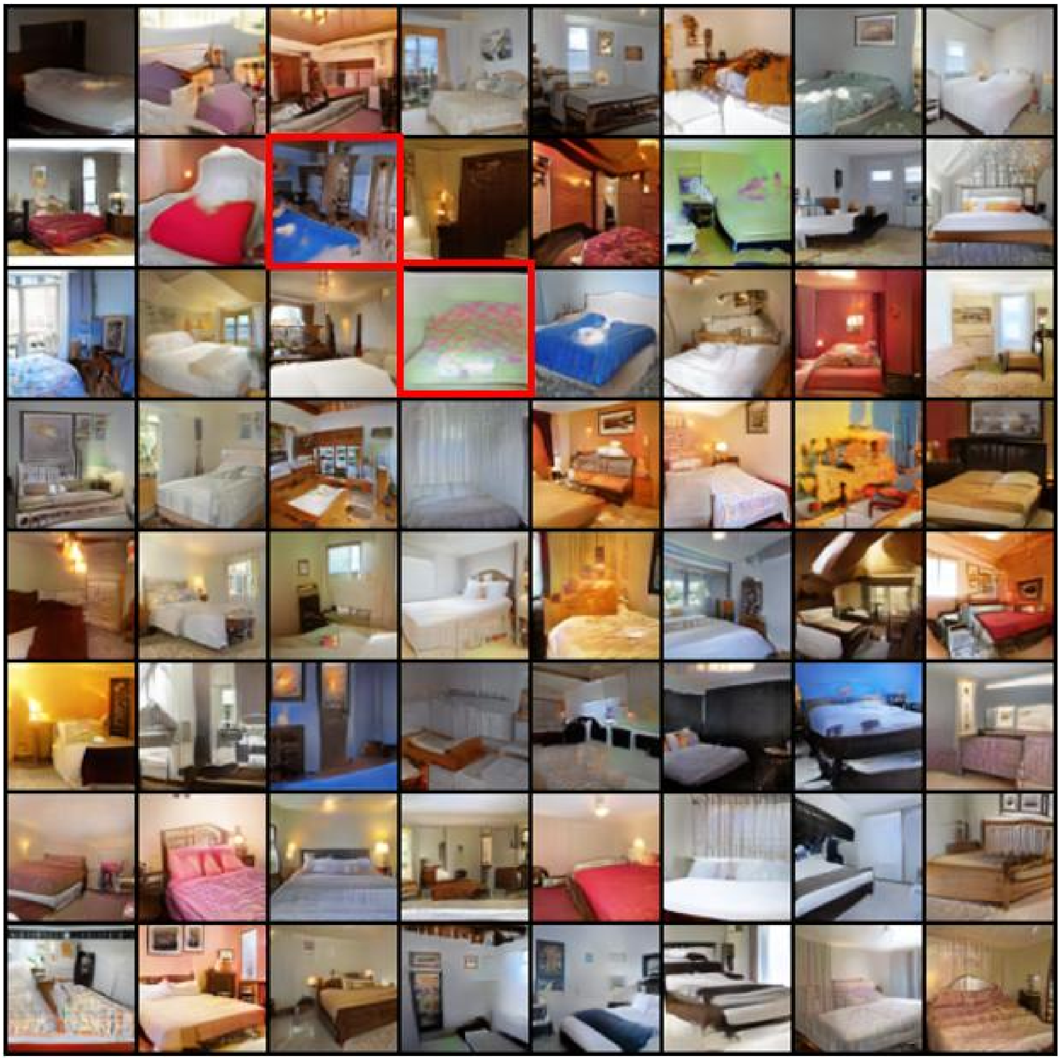

Figure 10. Larger version of Figure 6 (d) in the main paper (WGAN-QC). Obvious failure cases are marked with red boxes.

References

[1] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, 2015.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations*, 2018.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.

[7] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the International Conference on Machine Learning*, 2018.

[8] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.

[9] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[10] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. Wasserstein divergence for GANs. In *Proceedings of the European Conference on Computer Vision*, 2018.

[11] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[12] Yasin Yazıcı, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The unusual effectiveness of averaging in GAN training. In *Proceedings of the International Conference on Learning Representations*, 2019.