ManiGAN: Text-Guided Image Manipulation

Bowen Li\textsuperscript{1} Xiaojuan Qi\textsuperscript{1,2} Thomas Lukasiewicz\textsuperscript{1} Philip H. S. Torr\textsuperscript{1}
\textsuperscript{1}University of Oxford \textsuperscript{2}University of Hong Kong
{bowen.li, thomas.lukasiewicz}@cs.ox.ac.uk \{xiaojuan.qi, philip.torr\}@eng.ox.ac.uk

A. Architecture

We adopt the ControlGAN\textsuperscript{3} as the basic framework and replace batch normalisation with instance normalisation\textsuperscript{6} everywhere in the generator network except in the first stage. Basically, the affine combination module (ACM) can be inserted anywhere in the generator, but we experimentally find that it is best to incorporate the module before up-sampling blocks and image generation networks; see Fig. 2.

A.1. Residual Block

Each residual block contains two convolutional layers, two instance normalisation (IN)\textsuperscript{6}, and one GLU\textsuperscript{1} non-linear function. The architecture of the residual block used in the detail correction module is shown in Fig. 1.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{residual_block.png}
\caption{The architecture of the residual block.}
\end{figure}

B. Objective Functions

We train the main module and detail correction module separately, and the generator and discriminator in both modules are trained alternatively by minimising both the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$.

**Generator objective.** The loss function for the generator follows those used in ControlGAN\textsuperscript{3}, but we introduce a regularisation term:

$$\mathcal{L}_{\text{rec}} = 1 - \frac{1}{CHW} ||I' - I||, \quad (1)$$

to prevent the network achieving identity mapping, which can penalise large perturbations when the generated image becomes the same as the input image.

$$\mathcal{L}_G = \frac{1}{2} E_{I' \sim P_{\text{data}}} [\log(D(I'))] - \frac{1}{2} E_{I' \sim P_{\text{data}}} [\log(1 - D(I'))]$$

$$+ \frac{1}{2} E_{I \sim P_{\text{data}}} [\log(1 - D(I))] - \frac{1}{2} E_{I' \sim P_{\text{data}}} [\log(D(I'))]$$

$$+ \frac{1}{2} E_{I \sim P_{\text{data}}} [\log(1 - D(I))]$$

$$+ \lambda_3 (1 - \mathcal{L}_{\text{core}}(I, S')) + \mathcal{L}_{\text{core}}(I', I), \quad (2)$$

where $I$ is the real image sampled from the true image distribution $P_{\text{data}}$, $S$ is the corresponding matched text that correctly describes the $I$, $I'$ is the generated image sampled from the model distribution $P_{\text{G}}$. The unconditional adversarial loss makes the synthetic image $I'$ indistinguishable from the real image $I$, the conditional adversarial loss aligns the generated image $I'$ with the given text description $S$, $\mathcal{L}_{\text{DAMSM}}$\textsuperscript{3} measures the text-image similarity at the word-level to provide fine-grained feedback for image generation, $\mathcal{L}_{\text{core}}$\textsuperscript{3} determines whether word-related visual attributes exist in the image, and $\mathcal{L}_{\text{rec}}$\textsuperscript{3} reduces randomness involved in the generation process. $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ are hyperparameters controlling the importance of additional losses. Note that we do not use $\mathcal{L}_{\text{rec}}$ when we train the detail correction module.

**Discriminator objective.** The loss function for the discriminator follows those used in ControlGAN\textsuperscript{3}, and the function used to train the discriminator in the detail correction module is the same as the one used in the last stage of the main module.

$$\mathcal{L}_D = \frac{1}{2} E_{I \sim P_{\text{data}}} [\log(D(I))] - \frac{1}{2} E_{I' \sim P_{\text{data}}} [\log(1 - D(I'))]$$

$$+ \frac{1}{2} E_{I \sim P_{\text{data}}} [\log(1 - D(I))] - \frac{1}{2} E_{I' \sim P_{\text{data}}} [\log(D(I'))]$$

$$+ \lambda_3 (1 - \mathcal{L}_{\text{core}}(I, S')) + \mathcal{L}_{\text{core}}(I', I), \quad (3)$$

where $S'$ is a given text description randomly sampled from the dataset. The unconditional adversarial loss determines whether the given image is real, and the conditional adversarial loss reflects the semantic similarity between images and texts.
C. Trend of Manipulation Results

We track the trend of manipulation results over epoch increases, as shown in Figs. 3 and 4. The original images are smoothly modified to achieve the best balance between the generation of new visual attributes (e.g., blue head, blue wings and yellow belly in Fig. 3) and the reconstruction of text-irrelevant contents of the original images (e.g., the shape of the bird and the background in Fig. 3) and the reconstruction of text-irrelevant contents of the original images (e.g., the shape of the bird and the background in Fig. 3) and the reconstruction of text-irrelevant contents of the original images (e.g., the shape of the bird and the background in Fig. 3). However, when the epoch goes larger, the generated visual attributes (e.g., blue head, blue wings, and yellow belly of the bird, dirt background of the zebras) aligned with the given text descriptions are gradually erased, and the synthetic images become more and more similar to the original images. This verifies the existence of the trade-off between the generation of new visual attributes required in the given text descriptions and the reconstruction of text-irrelevant contents existing in the original images.

D. Additional Comparison Results

In Figs. 5, 6, 7, and 8 we show additional comparison results between our ManiGAN, SISGAN, and TAGAN on the CUB and COCO datasets. Please watch the accompanying video for detailed comparison.
This bird is **blue** and **grey** with a **red belly**.

This bird has wings that are **grey** and **yellow** with a **yellow belly**.

This bird is **black** in colour, with a **red crown** and a **red beak**.

This green bird has a **black crown** and a **green belly**.

A bird with **brown wings** and a **yellow body**, with a **yellow head**.

A white bird with **grey wings** and a **red bill**, with a **white belly**.

---

**Figure 5.** Additional comparison results between ManiGAN, SISGAN, and TAGAN on the CUB bird dataset.
A small blue bird with an orange crown, with a grey belly.

This bird has a red head, black eye rings, and a yellow belly.

This bird is mostly red with a black beak, and a black tail.

This tiny bird is blue and has a red bill and a red belly.

This bird has a white head, a yellow bill, and a yellow belly.

A white bird with red throat, black eye rings, and grey wings.

Given Text  | Original | SISGAN | TAGAN | Ours
---|---|---|---|---
Figure 6. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the CUB bird dataset.
Sunset.

Blue boat, green grass.

White bus.

Man, dry grass.

Brown cow, dirt.

Boy, road.

Given Text

Original

SISGAN

TAGAN

Ours

Figure 7. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the COCO dataset.
Zebra, grass.

Orange bus.

Night.

Kite, green field.

Pizza, pepperoni.

Zebra, water.

Figure 8. Additional comparison results between ManiGAN, SISGAN, and TAGAN on the COCO dataset.
References


