# Generative Semantic Manipulation with Mask-Contrasting GAN

Xiaodan Liang[1(✉)], Hao Zhang[1], Liang Lin[2], and Eric Xing[1]

[1] Carnegie Mellon University, Pittsburgh, USA
{xiaodan1,hao,epxing}@cs.cmu.edu
[2] Sun Yat-sen University, Guangzhou, China
linliang@ieee.org

**Abstract.** Despite the promising results on paired/unpaired image-to-image translation achieved by Generative Adversarial Networks (GANs), prior works often only transfer the low-level information (e.g. color or texture changes), but fail to manipulate high-level semantic meanings (e.g., geometric structure or content) of different object regions. On the other hand, while some researches can synthesize compelling real-world images given a class label or caption, they cannot condition on arbitrary shapes or structures, which largely limits their application scenarios and interpretive capability of model results. In this work, we focus on a more challenging semantic manipulation task, aiming at modifying the semantic meaning of an object while preserving its own characteristics (e.g. viewpoints and shapes), such as cow→sheep, motor→bicycle, cat→dog. To tackle such large semantic changes, we introduce a contrasting GAN (contrast-GAN) with a novel adversarial contrasting objective which is able to perform all types of semantic translations with one category-conditional generator. Instead of directly making the synthesized samples close to target data as previous GANs did, our adversarial contrasting objective optimizes over the distance comparisons between samples, that is, enforcing the manipulated data be semantically closer to the real data with target category than the input data. Equipped with the new contrasting objective, a novel mask-conditional contrast-GAN architecture is proposed to enable disentangle image background with object semantic changes. Extensive qualitative and quantitative experiments on several semantic manipulation tasks on ImageNet and MSCOCO dataset show considerable performance gain by our contrast-GAN over other conditional GANs.

**Keywords:** Generative Adversarial Network
Image semantic manipulation

## 1 Introduction

Arbitrarily manipulating image content given either a target image, class or caption has recently attracted a lot of research interests and would advance a


V. Ferrari et al. (Eds.): ECCV 2018, LNCS 11217, pp. 574–590, 2018.
https://doi.org/10.1007/978-3-030-01261-8_34

wide range of applications, e.g. image editing and unsupervised representation learning. Recent generative models [4,13,15,17,36,42,43,46] have achieved great progress on modifying low-level content, such as transferring color and texture from a holistic view. However, these models often tend to ignore distinct semantic information (e.g. background or objects) conveyed at distinct image regions and directly render the whole image with one holistic color/texture. This largely limits the application potential of image generation/translation tasks where large semantic changes (e.g. cat → dog, motor → bicycle) are more appealing and essential to bridge the gap between high-level concepts and low-level image processing.


**Fig. 1.** Some example semantic manipulation results by our model, which takes one image and the desired object category (e.g. *cat*, *dog*) as inputs and then learns to automatically change the object semantics by modifying their appearance or geometric structure. We show the original image (left) and manipulated result (right) in each pair.

On the other hand, compelling conditional image synthesis given a specific object category (e.g. "bird") [29,41], a textual description ("a yellow bird with a black head") [33], or locations [34] has already been demonstrated using variants of Generative Adversarial Networks (GANs) [9,32] and Variational Autoencoders [10]. However, existing approaches have so far only used fixed and simple conditioning variables such as a class or location that can be conveniently formatted as inputs, but failed to control more complex variables (e.g. shapes and viewpoints). It is thus desirable to endow the unsupervised generation models with the interpretive and controllable capability.

In this paper, we take a further step towards image semantic manipulation in the absence of any paired training examples. It not only generalizes

image-to-image translation research by enabling manipulate high-level object semantics but also pushes the boundary of controllable image generation research by retaining intrinsic characteristics conveyed in the original image as much as possible. Figure 1 shows some example semantic manipulation results by our model. Our model can successfully change the semantic meaning of the objects into desired ones, such as cat→dog by manipulating the original shape, geometric or texture of objects in the original image. Note that our model often manipulates the object shapes and structures of target regions to make them more likely be the target semantic.

To tackle such large semantic changes, we propose a novel contrasting GAN (contrast-GAN) in the spirit of learning by comparisons [11,37]. Different from the objectives used in previous GANs that often directly compare the target values with the network outputs, the proposed contrast-GAN introduces an adversarial distance comparison objective for optimizing one conditional generator and several semantic-aware discriminators. This contrasting objective enforces that the features of synthesized samples are much closer to those of real data with the target semantic than the input data. In addition, distinguished from existing GANs [8,26,45,46] that require training distinct generators and discriminators for each type of semantic/style translation, our contrast-GAN only needs to train one single conditional generator for all types of semantic translations benefiting from the category-conditional network structure.

In order to transform object semantics while keeping their original charismatics as much as possible (e.g. only manipulating animal faces for translating dog to cat), exploiting the distinct characteristics that depict different object semantics is thus very critical. Distinguished from the commonly used ranking loss, the new contrasting objective has two merits: (a) the approximated feature center by considering a set of randomly selected instances with target semantic, can statistically learn the crucial characteristics determining each semantic; (b) The competition between two distance pairs of the desired object with original features and approximated feature center of target semantic enables to learn a good balance between semantic manipulation and characteristic-preserving, leading to a controllable system. Compared to simple object replacement, the controllable manipulation is critical for some applications (e.g. image editing). Such competition objective also alleviates the model collapse into average object appearance, like that other GANs suffer from.

In order to disentangle image background from semantic object regions, we further propose a novel mask-conditional contrast-GAN architecture for realizing the attentive semantic manipulation on the whole image by conditioning on masks of object instances. A category-aware local discriminator is employed to examine the fidelity and manipulated semantics of generated object regions, while a whole-image discriminator is responsible for the appearance consistency of the manipulated object regions and image backgrounds. Note that our model is general for taking any mask resources as inputs for each image, such as human specified masks or mask results by any segmentation methods [2,20–23,27,40].

We demonstrate the promising semantic manipulation capability of the proposed contrast-GAN on labels↔photos on Cityscape dataset [3], apple↔orange and horse↔zebra on Imagenet [5] and ten challenging semantic manipulation tasks (e.g. cat↔dog, bicycle↔motorcycle) on MSCOCO dataset [24], as illustrated in Fig. 1. We further quantitatively show its superiority compared to existing GAN models [8,15,26,38,46] on unpaired image-to-image translation task and more challenging semantic manipulation tasks.

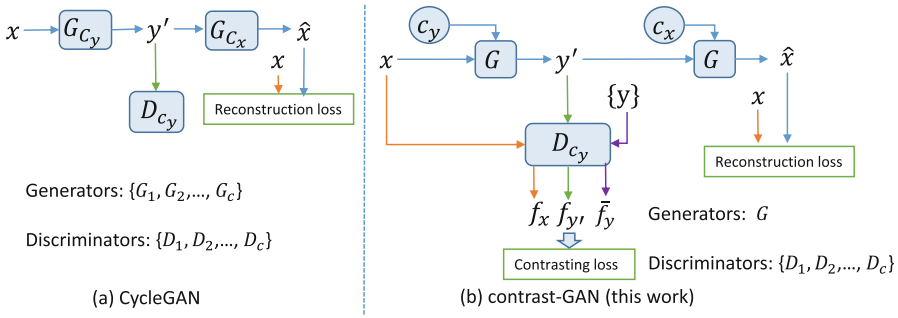


**Fig. 2.** An overview of the proposed contrast-GAN. $c_y$ and $c_x$ indicate the object categories (semantics) of domain $X$ and $Y$, respectively. $G_{c_y}$ translates samples into domain $Y$ and $D_{c_y}$ distinguishes between the manipulated result $y'$ and real ones $y$, and vice verse for $G_{c_x}$ and $D_{c_x}$. (a) shows the original CycleGAN in [46] where separate generators and discriminators for each mapping are optimized using the cycle-consistency loss. (b) presents the workflow of our contrast-GAN that optimizes one conditional generator $G$ and several semantic-aware discriminators $D_1, D_2, \ldots, D_C$, where $C$ is the total number of object categories. We introduce an adversarial contrasting loss into GAN that encourages the features $f_{y'}$ of generated sample $y'$ are much closer to the feature center $\bar{f}_y$ of target domain $Y$ than those of input $x$.

## 2 Related Work

**Generative Adversarial Networks (GANs).** There have been ever-growing GAN-family methods since the seminal work by Goodfellow et al. [9]. Impressive progresses have been achieved on a wide variety of image generation [6,7,29,34,35,44], image editing [45], text generation [12,18] and conditional image generation such as text2image [33], image inpainting [30], and image translation [13,19] tasks. The key to GANs' success is the variants of adversarial loss that forces the synthesized images to be indistinguishable from real data distribution. To handle the well-known mode collapse issue of GAN and make its training more stable, diverse training objectives have been developed, such as Earth Mover Distance in WGAN [1], feature matching loss [35], loss-sensitive GAN [31]. However, unlike existing GAN objectives that seek an appropriate criterion between synthesized samples and target outputs, we propose a tailored adversarial contrasting objective for image semantic manipulation. Our contrast-GAN is inspired

by the strategy of learning by comparison, that is, aiming to learn the mapping function such that the semantic features of manipulated images are much closer to feature distributions of target domain than those of the original domain.

**Generative Image-conditional Models.** GANs have shown great success on a variety of image-conditional models such as style transfer [15,39] and general-purpose image-to-image translation [13]. More recent approaches [25,26,43,46] have tackled the unpaired setting for cross-domain image translation and also conducted experiments on simple semantic translation (e.g. horse→zebra and apple→orange), where only color and texture changes are required. Compared to prior approaches that only transfer low-level information, we focus on high-level semantic manipulation on images given the desired category. The unified mask-controllable contrast-GAN is introduced to disentangle image background with object parts, comprised by one shared conditional generator and several semantic-aware discriminators within an adversarial optimization. Our model can be posed as a general-purpose solution for high-level semantic manipulation, which can facilitate many image understanding task, such as unsupervised and semi-supervised activity recognition and object recognition. Inspired by the dual-GAN [43] and Cycle-GAN [46] that learns the inverse mapping to constrain the network outputs, we also incorporate the cycle-consistency loss into our contrast-GAN architecture.

## 3 Semantic Manipulation with Contrasting GAN

The goal of semantic manipulation is to learn mapping functions for manipulating input images into target domains specified by various object semantics $\{c_k\}_{k=1}^{C}$, where $C$ is the total number of target categories. For each semantic $c_k$, we have a set of images $\{I_{c_k}\}$. For notation simplicity, we denote the input domain as $X$ with semantic $c_x$ and output domain as $Y$ with semantic $c_y$ in each training/testing step. As illustrated in Fig. 2, our contrast-GAN learns a conditional generator $G$, which takes a desired semantic $c_y$ and an input image $x$ as inputs, and then manipulates $x$ into $y'$. The semantic-aware adversarial discriminators $D_{c_y}$ aims to distinguish between images $y \in Y$ and manipulated results $y' = G(x, c_y)$. Our new adversarial contrasting loss forces the representations of $y'$ be closer to those of images $\{y\}$ in target domain $Y$ than those of input image $x$.

In the following sections, we first describe our contrast-GAN architecture and then present the mask-conditional contrast-GAN for disentangling image background and object semantics.

### 3.1 Adversarial Contrasting Objective

The adversarial loss introduced in Generative Adversarial Networks (GANs) [9] consists of a generator $G$ and a discriminator $D$ that compete in a two-player min-max game. The objective of vanilla GAN is to make the discriminator correctly classify its inputs as either real or synthetic and the generator synthesize

images that the discriminator will classify as real. In practice, we can replace the negative log-likelihood objective by a least square loss [28], which performs more stable during training and generates higher quality results. Thus, the GAN objective becomes:

$$\begin{aligned}\mathcal{L}_{\text{LSGAN}}(G, D_{c_y}, c_y) = \mathbb{E}_{y\sim p_{\text{data}}(y)}[(D_{c_y}(y)-1)^2] \\ + \mathbb{E}_{x\sim p_{\text{data}}(x)}[D_{c_y}(G(x, c_y))^2].\end{aligned} \tag{1}$$

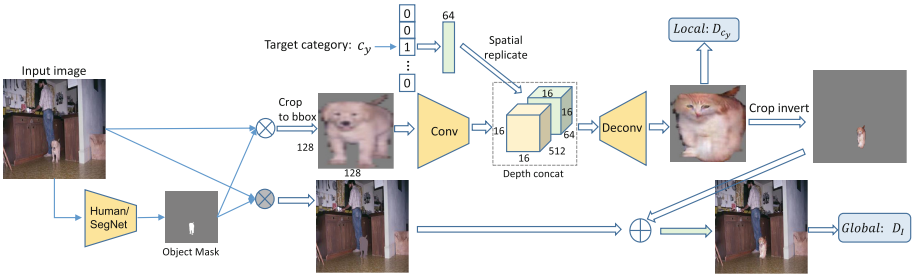


**Fig. 3.** The proposed mask-conditional contrast-GAN for semantic manipulation by taking an input image, an object mask and a target category as input. Any segmentation methods can be used to produce the object masks with input categories.

In this work, in order to tackle large semantic changes, we propose a new adversarial contrasting objective in the spirit of learning by comparison. Using a comparative measure of the neural network to learn embedding space was introduced in the "Siamese network" [11,37] with triple samples. The main idea is to optimize over distance comparisons between generated samples and those from the source domain $X$ and target domain $Y$. We consider the feature representation of manipulated result $y'$ should be closer to those of real data $\{y\}$ in target domain $Y$ than that of $x$ in input domain $X$ under the background of object semantic $c_y$. Formally, we can produce semantic-aware features by feeding the samples into $D_{c_y}$, resulting in $f_{y'}$ for $y'$ served as an anchor sample, $f_x$ for the input $x$ as a contrasting sample and $\{f_y\}_N$ for samples $\{y\}_N$ in the target domain as positive samples. Note that, at each training step, we compare the anchor $f_{y'}$ with the approximated feature center $\bar{f}_y$ computed as the average of all features $\{f_y\}_N$ rather than that of one randomly sampled $y$ in each step, in order to reduce model oscillation. The generator aims to minimize the contrasting distance $Q(\cdot)$:

$$Q(f_{y'}, f_x, \bar{f}_y) = -\log \frac{e^{-||f_{y'}-\bar{f}_y||_2}}{e^{-||f_{y'}-\bar{f}_y||_2} + e^{-||f_{y'}-f_x||_2}}. \tag{2}$$

Similar to the target of $D_{c_y(y)}$ in Eq.(1) that tries to correctly classify its inputs as either real or fake, our discriminator aims to maximize the contrasting

distance $Q(f_{y'}, f_x, \bar{f}_y)$. The adversarial contrasting objective for GAN can be defined as:

$$\begin{aligned}\mathcal{L}_{\text{contrast}}(G, D_{c_y}, c_y) = \mathbb{E}_{y\sim p_{\text{data}}(y), x\sim p_{\text{data}(x)}} \\ [Q(D_{c_y}(G(x, c_y)), D_{c_y}(x), D_{c_y}(\{y\}))].\end{aligned} \tag{3}$$

To further reduce the space of possible mapping functions by the conditional generator, we also use the cycle-consistency loss in [46] which constrains the mappings (induced by the generator $G$) between two object semantics should be inverses of each other. Notably, different from [46] which used independent generators for each domain, we use a single shared conditional generator for all domains. The cycle objective can be defined as:

$$\mathcal{L}_{\text{cycle}}(G, c_y, c_x) = \mathbb{E}_{x\sim p_{\text{data}(x)}}[||G(G(x, c_y), c_x) - x||_1]. \tag{4}$$

Therefore, our full objective is computed by combining Eqs. (1), (3) and (4):

$$\begin{aligned}\mathcal{L}_{\text{contrast-GAN}}(G, D_{c_y}, c_y) = \mathcal{L}_{\text{contrast}}(G, D_{c_y}, c_y) \\ +\lambda\mathcal{L}_{\text{LSGAN}}(G, D_{c_y}, c_y) + \beta\mathcal{L}_{\text{cycle}}(G, c_y, c_x),\end{aligned} \tag{5}$$

where $\lambda$ and $\beta$ control the relative importance of the objectives. $G$ tries to minimize this objective against a set of adversarial discriminators $\{D_{c_y}\}$ that try to maximize them, i.e. $G^* = \arg\min_G(\frac{1}{C}\sum_{c_y}\max_{D_{c_y}}\mathcal{L}_{\text{contrast-GAN}}(G, D_{c_y}, c_y))$. Our extensive experiments show that each of objectives plays a critical role in arriving at high-quality manipulation results.

### 3.2 Mask-Conditional Contrast-GAN

Figure 3 shows a sketch of our model, which starts from an input image $x$, an object mask $M$ and target category $c_y$ and outputs the manipulated image. Note that the whole model is fully differential for back-propagation. For clarity, the full cycle architecture (i.e. the mapping $y' \rightarrow \hat{x}$ via $G(y, c_x)$) is omitted. Below we walk through each step.

First, a masking operation and subsequent spatial cropping operation are performed to obtain the object region with the size of $128\times128$. The background image is calculated by functioning the inverse mask map on an input image. The object region is then fed into several convolutional layers to get $16 \times 16$ feature maps with 512 dimension. Second, we represent the target category $c_y$ using a one-hot vector which is then passed into a linear layer to get a feature embedding with 64 dimension. This feature is replicated spatially to form a $16 \times 16 \times 64$ feature maps, and then concatenated with image feature maps via the depth concatenation. Third, several deconvolution layers are employed to obtain target region with $128 \times 128$. We then wrap the manipulated region back into the original image resolution, which is then combined with the background image via an additive operation to get the final manipulated image. We implement the spatial masking and cropping modules using spatial transformers [14].

To enforce the semantic manipulation results be semantically consistent with both the target semantic and the background appearance of the input image,

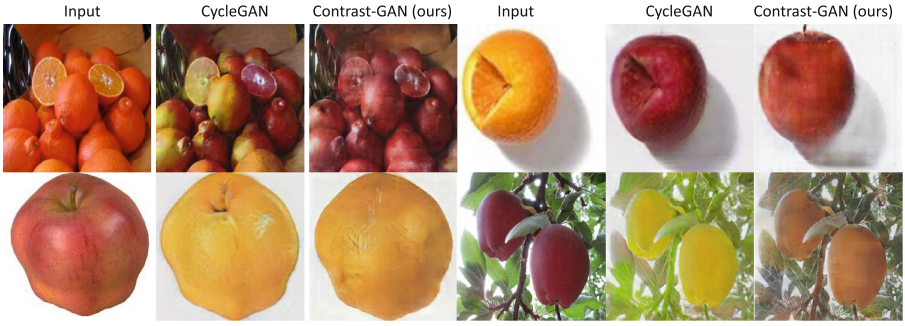


**Fig. 4.** Result comparison by our contrast-GAN with CycleGAN [46] for translating orange→apple (first row) and apple→orange (second row) on ImageNet, which demonstrates the advantage of leveraging adversarial contrasting objective in GAN.

we adopt both local discriminators $\{D_{c_y}\}$ defined in our contrast-GAN and a global image discriminator $D_I$. Each local discriminator $D_{c_y}$ is responsible for verifying whether the high-level semantic of outputs is semantically coherent with the input target while the global one $D_I$ evaluates the visual fidelity of the whole manipulated image. The global discriminator $D_I$ takes the combined image of the transformed regions and background, and randomly sampled image as inputs, then employ the same patch-level network as local discriminator, which is jointly trained with local discriminators.

### 3.3 Implementation Details

**Network Architecture.** To make a fair comparison, We adopt similar architectures from [46] which have shown impressive results for unpaired image translation. This generator contains three stride-2 convolutions, six residual blocks, and three fractionally stridden convolutions. For the architecture of mask-conditional contrast-GAN in Fig. 3, the residual blocks are employed after concatenating convolutional feature maps with maps of the target category. In terms of the target category input for generator $G$, we specify a different number of categories $C$ for each dataset, such as $C = 10$ for ten semantic manipulation tasks on MSCOCO dataset. We use the same patch-level discriminator used in [46] for local discriminators $\{D_{c_y}\}$ and the global discriminator $D_I$. By using the patch-level discriminator network, f(y) is a vector with 14*14 dimension.

**Training Details.** To compute the approximate feature center $\bar{f}_y$ in Eq.(2) for the contrasting objective, we keep an image buffer with randomly selected $N = 50$ samples in target domain $Y$. For all the experiments, we set $\lambda = 10$ and $\beta = 10$ in Eq. (5) to balance each objective. We use the Adam solver [16] with a batch size of 1. All networks were trained from scratch and trained with a learning rate of 0.0002 for the first 100 epochs and a linearly decaying rate that goes to zero over the next 100 epochs. Our algorithm only optimizes over one

conditional generator and several semantic-aware discriminators for all kinds of object semantics. All models are implemented on Torch framework.

**Table 1.** Comparison of FCN-scores on Cityscapes labels→photos.

| Method | Per-pixel acc. | Per-class acc. | Class IOU |
|---|---|---|---|
| CoGAN [26] | 0.40 | 0.10 | 0.06 |
| BiGAN [8] | 0.19 | 0.06 | 0.02 |
| Pixel loss + GAN [38] | 0.20 | 0.10 | 0.0 |
| Feature loss + GAN [15] | 0.07 | 0.04 | 0.01 |
| CycleGAN [46] | 0.52 | 0.17 | 0.11 |
| Contrast alone | 0.53 | 0.13 | 0.12 |
| Contrast + classify | 0.55 | 0.15 | 0.11 |
| Contrast + Cycle | 0.57 | 0.22 | 0.13 |
| Contrast-GAN (separate G) | 0.57 | **0.22** | **0.17** |
| **Contrast-GAN (ours)** | **0.58** | 0.21 | 0.16 |

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** First, we quantitatively compare the proposed contrast-GAN against recent state-of-the-arts on the task of labels↔photos on the Cityscape dataset [3]. The labels↔Photos dataset uses images from Cityscape training set for training and validation set for testing. Following [46], we use the unpaired setting during training and the ground truth input-output pairs for evaluation. Second, we compare our contrast-GAN with CycleGAN [46] on unpaired translation, evaluating on the task of horse↔zebra and apple↔orange from ImageNet. The images for each class are downloaded from ImageNet [5] and scaled to 128×128, consisting of 939 images for the horse, 1177 for zebra, 996 for apple and 1020 for orange. Finally, we apply contrast-GAN into ten more challenging semantic manipulation tasks, i.e. dog↔cat, cow↔sheep, bicycle↔motorcycle, horse↔giraffe, horse↔zebra. To disentangle image background with the object semantic information, we test the performance of mask-conditional architecture. The mask annotations for each image are obtained from MSCOCO dataset [24]. For each object category, the images in MSCOCO train set are used for training and those in MSCOCO validation set for testing. The output realism of manipulated results by different methods is quantitatively compared by AMT perception studies described below.

**Evaluation Metrics.** We adopt the "FCN score" from [13] to evaluate Cityscapes labels→photo task, which evaluates how interpretable the generated photos are according to an off-the-shelf semantic segmentation algorithm.

**Table 2.** Comparison of classification performance on Cityscapes photos→ labels dataset.

| Method | Per-pixel acc | Per-class acc | Class IOU |
|---|---|---|---|
| CoGAN [26] | 0.45 | 0.11 | 0.08 |
| BiGAN [8] | 0.41 | 0.13 | 0.07 |
| Pixel loss + GAN [38] | 0.47 | 0.11 | 0.07 |
| Feature loss + GAN [15] | 0.50 | 0.10 | 0.06 |
| CycleGAN [46] | 0.58 | 0.22 | 0.16 |
| Contrast alone Contrast + classify | 0.55 | 0.13 | 0.11 |
| Contrast + Cycle | 0.60 | 0.19 | 0.15 |
| Contrast-GAN (separate G) | 0.60 | 0.23 | 0.17 |
| **Contrast-GAN (ours)** | **0.61** | **0.23** | **0.18** |

To evaluate the performance of photo→labels, we use the standard "semantic segmentation metrics" from Cityscapes benchmark, including per-pixel accuracy, per-class accuracy, and mean class Intersection-Over-Union [3]. For semantic manipulation tasks on ImageNet and MSCOCO datasets (e.g. cat→dog), we run real vs.fake AMT perceptual studies to compare the realism of outputs from different methods under the background of a specific object semantic (e.g. dog), similar to [46]. For each semantic manipulation task, we collect 10 annotations for randomly selected 100 manipulated images by each method and all methods perform manipulation results on the same set of images.

### 4.2 Result Comparisons

**Labels↔photos on Cityscape.** Tables 1 and 2 report the performance comparison on the labels→photos task and photos→labels task on Cityscape, respectively. In both cases, the proposed contrast-GAN with a new adversarial contrasting objective outperforms the state-of-the-arts [8,15,26,38,46] on unpaired image-to-image translation. Note that we adopt the same baselines [8,15,26,38] for fair comparison in [46].

**Apple ↔orange and horse↔zebra on ImageNet.** Figure 4 shows some example results by the baseline CycleGAN [46] and our contrast-GAN on the apple↔orange semantic manipulation. It can be seen that our method successfully transforms the semantic of objects while CycleGAN only tends to modify low-level characteristics (e.g. color and texture). We also perform real vs. fake AMT perceptual studies on both apple↔orange and horse↔zebra tasks. Our contrast-GAN can fool participants much better than CycleGAN [46] by comparing the number of manipulated images that Turkers labeled real, that is 14.3% vs 12.8% on average for apple↔orange and 10.9% vs 9.6% on average for horse↔zebra.

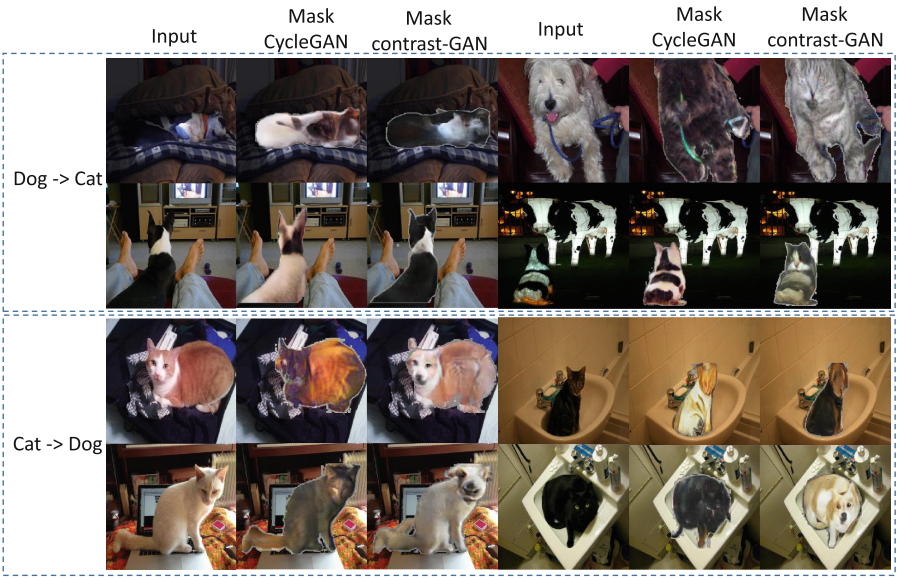


**Fig. 5.** Result comparison between our mask contrast-GAN with mask CycleGAN [46] for translating dog→cat and cat→dog on the MSCOCO dataset with provided object masks. It shows the superiority of adversarial contrasting objectiveness over the objectiveness used in CycleGAN [46].

### 4.3 Semantic Manipulation Tasks on MSCOCO

We further demonstrate the effectiveness of our method on ten challenging semantic manipulation applications with large semantic changes.

**Contrasting objective vs. other GAN objectives.** Figure 5 visualizes the comparisons of our mask-conditional architecture using cycle-consistency loss in [46] and our contrasting objective, that is, mask CycleGAN vs mask contrast-GAN. The baseline method often tries to translate very low-level information (e.g. color changes) and fails to edit the shapes and key characteristic (e.g. structure) that truly convey a specific high-level object semantic. However, our contrast-GAN tends to perform trivial yet critical changes on object shapes and textures to satisfy the target semantic while preserving the original object characteristics. In Table 3, we report quantitative comparison results with the state-of-the-art CoGAN [26], BiGAN [8] and CycleGAN [46] on the AMT perceptual realism measure for eight semantic manipulation tasks. It can be observed that our method substantially outperforms the baseline on all tasks, especially on those requiring large semantic changes (e.g. cat↔dog and bicycle↔motorcycle). In Fig. 6, we show more qualitative results. Our model shows the promising capability of manipulating object semantics while retaining original shapes, viewpoints, and interactions with the background.


**Fig. 6.** Example results by our mask contrast-GAN for manipulating a variety of object semantics on MSCOCO dataset. For each image pair, we show the original image (left) and manipulated image (right) by specifying a desirable object semantic.

### 4.4 The Effectiveness of Mask-Conditional Architecture

As observed from Fig. 7, the original GAN networks often renders the whole image with the target texture and ignores the particular image content at different locations/regions. It may result in wrongly translating the unrelated objects (e.g. person, building) and background as the stripe texture in the horse→zebra case. On the contrary, our mask-conditional framework shows appealing results with the capability of selectively manipulating objects of interest (e.g. horse) into the desired semantic (e.g. zebra). It should be noted that our mask-conditional is general enough that can support any mask resources, e.g. human-provided masks and segmentation regions produced by any segmentation methods [2].

### 4.5 The Effectiveness of Each Objective

In Tables 1 and 2, we report the results by different variants of our full model on Cityscape labels↔photos task. "Contrast alone" indicates the model only uses $\mathcal{L}_{\text{contrast}}$ as the final objective in Eq.(5) while "Contrast + classify" represents the usage of combining of $\mathcal{L}_{\text{contrast}}$ and $\mathcal{L}_{\text{LSGAN}}$ as the final objective. "Contrast + cycle" is the variant that removes $\mathcal{L}_{\text{LSGAN}}$. CycleGAN [46] can also be regarded as one simplified version of our model that removes the contrasting objective. Table 3 shows the ablation studies on mask-conditional semantic manipulation tasks on MSCOCO dataset. It can be seen that "Contrast alone" and "Mask

**Table 3.** Result comparison of AMT perception test on eight mask-conditional semantic-manipulation tasks on the MSCOCO dataset. The numbers indicate % images that Turkers labeled real.

| Method | cat→dog | dog→cat | bicycle→motor | motor→bicycle |
|---|---|---|---|---|
| Mask CoGAN [26] | 1.1% | 2.0% | 7.6% | 12.1% |
| Mask BiGAN [8] | 1.9% | 2.1% | 8.2% | 11.4% |
| Mask CycleGAN [46] | 2.5% | 4.1% | 10.9% | 15.6% |
| Mask Contrast alone | 3.7% | 5.0% | 9.3% | 13.1% |
| Mask Contrast-GAN w/o $D_I$ | 4.3% | 6.0% | 12.8% | 15.7% |
| Mask Contrast-GAN (gt) | **4.8%** | 6.2% | 13.0% | **16.7%** |
| **Mask Contrast-GAN (predict)** | 4.5% | **6.5%** | **13.1%** | 15.8% |
| Method | horse→ giraffe | giraffe→ horse | cow→sheep | sheep→cow |
| Mask CoGAN [26] | 0.1% | 0.9% | 11.2% | 15.3% |
| Mask BiGAN [8] | 1.2% | 1.5% | 12.5% | 16.8% |
| Mask CycleGAN [46] | 1.5% | 2.3% | 16.3% | 18.9% |
| Mask Contrast alone | 1.6% | 1.8% | 17.1% | 15.5% |
| Mask Contrast-GAN w/o $D_I$ | 1.9% | 4.5% | 18.3% | 19.1% |
| Mask Contrast-GAN (gt) | **1.9%** | 5.4% | 18.7% | 20.5% |
| **Mask Contrast-GAN (predict)** | 1.7% | **6.3%** | **18.9%** | **21.6%** |

Contrast alone" achieve comparable results with the state-of-the-arts. Removing the original *classification*-like objective $\mathcal{L}_{\text{LSGAN}}$ degrades results compared to our full model, as does removing the cycle-consistency objective $\mathcal{L}_{\text{Cycle}}$. Therefore, we can conclude that all three objectives are critical for performing the semantic manipulation. $\mathcal{L}_{\text{LSGAN}}$ can be complementary with our contrasting objective $\mathcal{L}_{\text{contrast}}$ on validating the visual fidelity of manipulated results. We also validate the advantage of using an auxiliary global discriminator $D_I$ by comparing "Mask Contrast-GAN w/o $D_I$" and our full model in Table 3.

### 4.6 One Conditional Generator vs. Separate Generators

Note that instead of using separate generators for each semantic as in previous works [8,15,26,38,46], we propose to employ a conditional generator shared for all object semantics. Using one conditional generator has two advantages: first, it can lead to more powerful and robust feature representation by learning over more diverse samples of different semantics; second, the model size can be effectively reduced by only feeding different target categories as inputs to achieve different semantic manipulations. Tables 1 and 2 also report the results of using separate generators for each semantic task in our model, that is, "Contrast-GAN (separate G)". We can see that our full model using only one conditional generator shows slightly better results than "Contrast-GAN (separate G)".

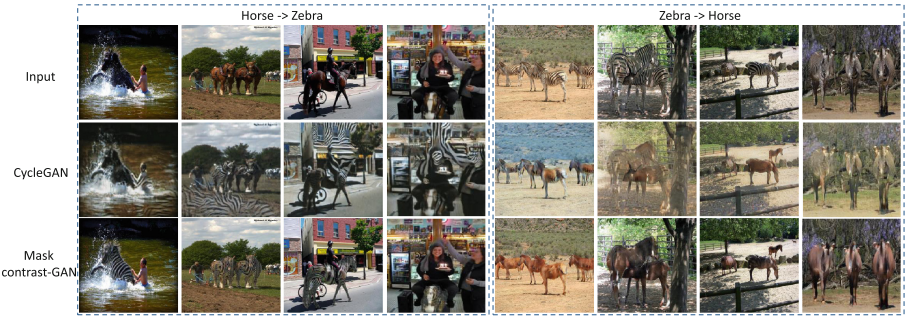


**Fig. 7.** Result comparisons between our mask contrast-GAN with CycleGAN [46] for translating horse→zebra and zebra→horse on the MSCOCO dataset. It shows the effectiveness of incorporating object masks to disentangle image background and object semantics.

### 4.7 The Effect of Different Mask Resources

Our mask-conditional architecture is able to manipulate any input images by firstly obtaining rough object masks of input categories using any segmentation methods [2], which is demonstrated by comparing "Mask Contrast-GAN (predict)" with "Mask Contrast-GAN (gt)" in Table 3. "Mask Contrast-GAN (predict)" indicates the results of using predicted masks by the segmentation model [2] as the network input. We can observe that no significant difference in visualization quality of manipulated images can be observed. The reason is that our model only needs a rough localization of objects with input categories and then generates the manipulated regions with new shapes and structures conditioned on the input object regions. Thus the inaccurate input masks will not significantly affect the manipulation performance.

## 5 Discussion and Future Work

This paper presents a novel adversarial contrasting objective and mask-conditional architecture, which together achieve compelling results in many semantic manipulation tasks. However, it still shows unsatisfactory results for some cases which require very large geometric changes, such as car↔truck and car↔bus. Integrating spatial transformation layers for explicitly learning pixel-wise offsets may help resolve very large geometric changes. To be more general, our model can be extended to automatically learned attentive regions via attention modeling. This paper pushes forward the research of unsupervised setting by demonstrating the possibility of manipulating high-level object semantics rather than the low-level color and texture changes as previous works did. In addition, it would be more interesting to develop techniques that are able to manipulate object interactions and activities in images/videos.

## References


1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. In: ICLR (2017)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv preprint arXiv:1606.00915 (2016)
3. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: CVPR, pp. 3213–3223 (2016)
4. Dai, W., et al.: Scan: structure correcting adversarial network for chest x-rays organ segmentation. arXiv preprint arXiv:1703.08770 (2017)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)
6. Deng, Z., et al.: Structured generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 3899–3909 (2017)
7. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. arXiv preprint arXiv:1605.09782 (2016)
8. Dumoulin, V., et al.: Adversarially learned inference. arXiv preprint arXiv:1606.00704 (2016)
9. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
10. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: a recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)
11. Hoffer, E., Hubara, I., Ailon, N.: Deep unsupervised learning through spatial contrasting. arXiv preprint arXiv:1610.00243 (2016)
12. Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., Xing, E.P.: Controllable text generation. arxiv preprint. arXiv preprint arXiv:1703.00955, p. 7 (2017)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016)
14. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS, pp. 2017–2025 (2015)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV, pp. 694–711 (2016)
16. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S.: Perceptual generative adversarial networks for small object detection. In: CVPR (2017)
18. Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent topic-transition GAN for visual paragraph generation. In: ICCV (2017)
19. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion GAN for future-flow embedded video prediction. In: IEEE International Conference on Computer Vision (ICCV), vol. 1 (2017)
20. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph LSTM. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 125–143. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_8
21. Liang, X., et al.: Reversible recursive instance-level object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641 (2016)
22. Liang, X., Zhou, H., Xing, E.: Dynamic-structured semantic propagation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 752–761 (2018)

23. Lin, L., Wang, G., Zhang, R., Zhang, R., Liang, X., Zuo, W.: Deep structured scene parsing by learning with image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2276–2284 (2016)
24. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. arXiv preprint arXiv:1703.00848 (2017)
26. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NIPS, pp. 469–477 (2016)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
28. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z.: Multi-class generative adversarial networks with the l2 loss function. arXiv preprint arXiv:1611.04076 (2016)
29. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
30. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: CVPR, pp. 2536–2544 (2016)
31. Qi, G.J.: Loss-sensitive generative adversarial networks on lipschitz densities. arXiv preprint arXiv:1701.06264 (2017)
32. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
33. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
34. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: NIPS, pp. 217–225 (2016)
35. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. arXiv preprint arXiv:1606.03498 (2016)
36. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: controlling deep image synthesis with sketch and color. In: CVPR (2017)
37. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: CVPR, pp. 815–823 (2015)
38. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
39. Wang, H., Liang, X., Zhang, H., Yeung, D.Y., Xing, E.P.: ZM-Net: real-time zero-shot image manipulation network. arXiv preprint arXiv:1703.07255 (2017)
40. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: IEEE CVPR, vol. 1, p. 3 (2017)
41. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2Image: conditional image generation from visual attributes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 776–791. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_47
42. Yang, L., Liang, X., Xing, E.: Unsupervised real-to-virtual domain unification for end-to-end highway driving. arXiv preprint arXiv:1801.03458 (2018)
43. Yi, Z., Zhang, H., Gong, P.T., et al.: DualGAN: unsupervised dual learning for image-to-image translation. In: ICCV (2017)

44. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.: Stack-Gan: text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
45. Zhu, J.-Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 597–613. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_36
46. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)