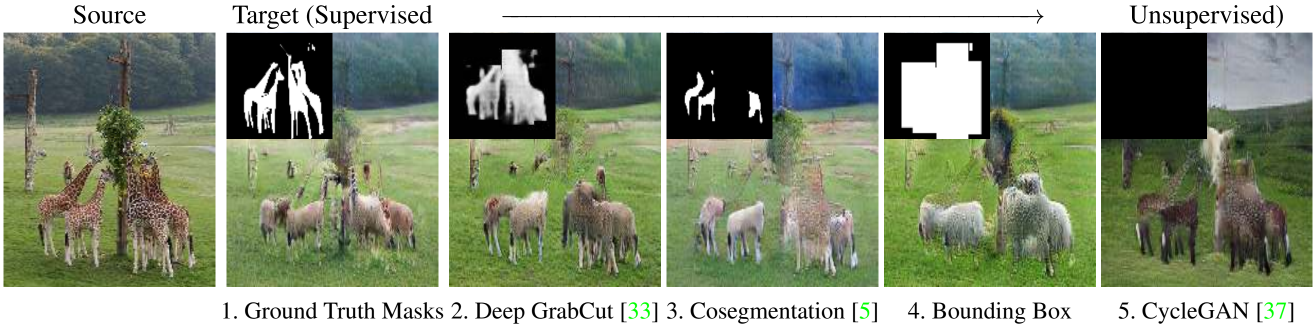
Exploring the Spectrum of Mask Supervision for Unpaired Image-to-Image Translation

Aaron Gokaslan

Brown University



Source

Target (Supervised → Unsupervised)

1. Ground Truth Masks 2. Deep GrabCut [33] 3. Cosegmentation [5] 4. Bounding Box 5. CycleGAN [37]

Figure 1: Results of varying supervision for the image-to-image object translation problem under significant shape change (giraffes→sheep). Less precise supervision is cheaper to acquire and still allows for acceptable image translation results.

Abstract

Current unsupervised image-to-image translation techniques have proven to struggle when faced with domains exhibiting significant shape changes. Recently proposed approaches enable this property but require instance-level segmentations which limit their domain applicability due to the cost of acquiring this supervision. We explore and analyze the effect of different forms of supervised masks for image-to-image translation under significant shape change. We show qualitatively and quantitatively that imprecise supervisions can produce appealing results, sometimes comparable to the full-supervision case.

1. Acknowledgements

I would like to thank Youssef Alami Mejjati, Oliver Wang, Prof. Kwang In Kim, and my advisor Prof. James Tompkin for all of their meaningful help and collaboration they have provided on this project. I could not have done it with all your help. I would also like to thank TStaff and Brown CS for providing the compute resources needed to run this experiment.

I ran most of the actual experiments for this research project while Youssef helped collect, analyze, and graph much of the results. Oliver Wang, Prof. Kwang In Kim, and Prof. James Tompkin all helped with paper editing, formatting, and providing valuable feedback.

2. Introduction

Data driven generative models like generative adversarial networks (GANs) [11] enable exciting new possibilities for image editing via powerful higher-level controls. We consider the image-to-image translation problem of mapping object appearance from one class to another when the objects have different shapes, *e.g.*, turning giraffes into sheep (Figure 1). Like most data-driven methods, the availability of training data is critical to the practical success of these approaches. Networks like InstaGAN [2] and ContrastGAN [21] require pixel-wise mask segmentation supervision but produce relatively high-quality results; however, such supervision may not be available and can be expensive to acquire, making unsupervised and imprecisely-supervised systems more appealing. In general, there is a spectrum of possible supervision, and discovering the 'sweet spot' would save time in supervision collection.

We conduct an experimental evaluation and analysis of image translation under shape change approaches across this spectrum of supervision:

1. Human-labeled pixel-wise masks (InstaGAN [2] style),
2. Ground-truth bounding boxes with automatically-refined boundaries (via Deep GrabCut [33] style),
3. Cosegmentation with another image of the same object (Chen et al. [5] style),
4. Ground-truth bounding boxes, and
5. No supervision (CycleGAN [37] style).

We evaluate a set of different loss terms with respect to these supervisions: relativistic discriminators, multi-scale discriminators, feature matching loss, cyclic mask loss, shared background context loss, and mask adversarial loss without RGB adversarial loss.

From this analysis, we present an improved architecture and loss to increase result quality over the state of the art, plus discuss approaches which did not mark a useful point on the spectrum (*e.g.*, unsupervised mask recovery (UA-GAN [25] style)). The take-away conclusion is that exact object masks are not required; we can generate reasonable quality results with automatically-refined bounding box supervision, and in certain object classes even with only an additional second image for cosegmentation.

Code and data used in our experiments will be made available to the research community.

3. Related Work

Many works in computer vision learn segmentation masks; as such, we will present a sample of current works on deep-learning-based unsupervised segmentation, and then present works on image-to-image translation with respect to segmentation supervision.

Unsupervised segmentation. Xia et al. [32] (W-Net) generate unsupervised instance-level image segmentations. A U-Net encoder produces a dense classification map, and a binary mask is computed with normalized cuts. Next, a U-Net decoder network reconstructs the input from the generated segmentation. Training occurs by minimizing a reconstruction loss, along with minimizing the difference between the classification and the normalized cut mask.

Cosegmentation. Given a set of images depicting the same object, the aim of cosegmentation is to find a mask per image covering the common content. Hsu et al. [13] recover cosegmentation masks in an unsupervised manner by enforcing for each pair of images that: 1) foreground feature representations are close with respect to the L2 norm, and 2) foreground and background feature representations in the same image should be distinct. Chen et al. [5] learn object cosegmentation in a supervised manner. To learn which feature map channels are useful across both images, they condition one image on a channel-wise weighting of each feature map from the other image. When trained on PASCAL VOC, this network generalizes well to unseen classes. We use Chen et al.'s approach to create sgementaton mask inputs, and consider it an imprecise form of supervision as it has not been trained on the datasets which we use (or even the object classes, except 'sheep').

Unpaired object transfiguration. Zhu et al. [37] (CycleGAN) and Kim et al. [18] (DiscoGAN) both demonstrated the ability to swap the texture of two objects with similar shape, such as transforming horses into zebras. These types of cyclic models can also translate between two domains of differing shapes and poses [10]. However, they struggle to deal with both small objects and domains that have different distributions of background texture.

Several recent papers [6, 23, 25, 34, 36] have proposed augmenting these networks with unsupervised attention mechanisms to help separate foreground and background regions. For instance, Yang et al. [35] (LR-GAN) decompose the regions with a two-component spatial-transformer generator: one component generates the background, and the other is a recurrent neural network (RNN) which repeatedly 'pastes' created foregrounds and their masks into the image. While promising, all these techniques struggle when transforming between domains of differing shape.

Liang et al. [21] (ContrastGAN) and Anonymous et al. [2] (InstaGAN) have proposed using binary mask segmentations to simplify the complex task of localizing objects and segmenting them from the background. While these approaches are effective, doing so requires a dataset of pixel-level segmentation which are expensive to label. This limits their use for novel categories of objects.

We propose a novel framework which builds upon the InstaGAN architecture, but is designed for mask-level supervision, rather than instance-level. We propose a number of new losses, which allow the network to use segmentations of varying accuracy as supervision. We demonstrate that even low quality segmentations can dramatically improve upon CycleGAN's unsupervised results.

4. Method

Suppose that we are given databases S and T of images of two object classes, each with varying object pose, object number, background, *etc*. sampled from the respective domains $\mathcal{S}$ and $\mathcal{T}$. Further, assume that each source

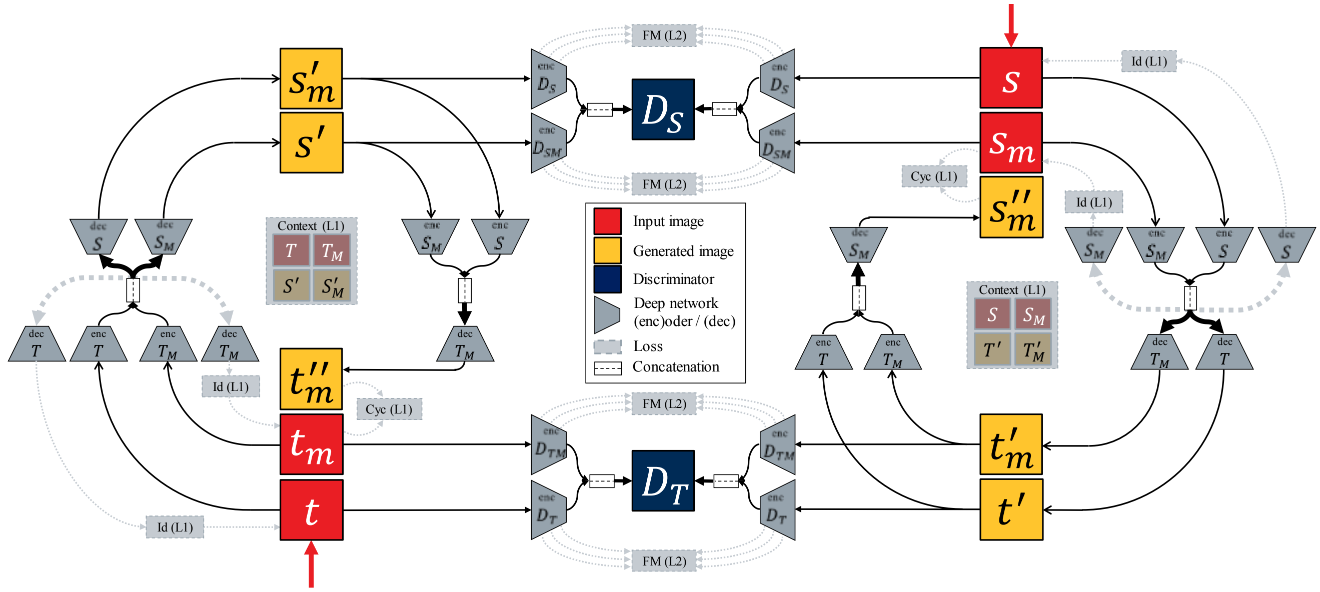


Figure 2: Architecture diagram for our system. The translation network cycles from domains $\mathcal{S}$ to $\mathcal{T}$ on the right-hand side, and from domains $\mathcal{T}$ to $\mathcal{S}$ on the left-hand side. Both domain discriminators D_s, D_t are mutually trained with the generators. All loss functions described in Section 4.1 are shown as light grey boxes. This architecture is highly motivated by InstaGAN [2], with added losses and without per-instance domain mapping.

image $\mathbf{s} \in \mathcal{S}$ is provided with an object segmentation mask $\mathbf{s_m} \in \mathcal{S}_\mathbf{m}$. Our goal is to learn a transfer function $f : \mathcal{S} \times \mathcal{S}_\mathbf{m} \rightarrow \mathcal{T}$ which translates objects in a source image $\mathbf{s}$ into objects in the target domain $\mathcal{T}$ in an output image $\mathbf{t}' := f(\mathbf{s}, \mathbf{s_m})$. Figure 2 visualizes our approach; please start at the red arrow in the top right.

We may obtain $\mathbf{t}'$ using an auto-encoder; however, we wish to incorporate information from the segmentation mask in the decoding process. As in InstaGAN [2], we encode both $\mathbf{s}$ and $\mathbf{s_m}$: $h_\mathbf{s} = \text{enc}_\mathcal{S}(\mathbf{s})$ and $h_{\mathbf{s}_m} = \text{enc}_{\mathcal{S}_\text{M}}(\mathbf{s_m})$. Then, we concatenate these feature representations channel-wise and feed them: 1) into a decoder $\text{dec}_\mathcal{T}$ which gives us $\mathbf{t}'$, and 2) into a mask decoder $\text{dec}_{\mathcal{T}_\text{M}}$ which gives us $\mathbf{t}'_\mathbf{m}$. Formally this is expressed via $\mathbf{t}' = G_\mathcal{S}(h_\mathbf{s}, h_{\mathbf{s_m}})$, and $\mathbf{t}'_\mathbf{m} = G_{\mathcal{S}_\text{M}}(h_\mathbf{s}, h_{\mathbf{m_s}})$.

To constrain our final results to look realistic, we use an adversarial discriminator $\mathcal{D}_\mathcal{S}$ which distinguishes between real pairs $(\mathbf{s}, \mathbf{s_m})$ and fake pairs $(\mathbf{s}', \mathbf{s}'_m)$, where $\mathbf{s}'$ is the generated image from $\mathbf{t} \in \mathcal{T}$ (likewise for $\mathbf{s}'_\mathbf{m}$). The RGB and mask inputs are fed to the discriminator by first separately encoding them: $d_\mathbf{s} = \text{enc}_{D_\mathcal{S}}(\mathbf{s})$ and $d_{\mathbf{s_m}} = \text{enc}_{D_{\mathcal{S}_\text{M}}}(\mathbf{s_m})$. Then, similar to the generation part, we concatenate these feature representations channel-wise and feed them into a discriminator: $D_\mathcal{S}(d_\mathbf{s}, d_{\mathbf{s_m}})$ gives us the probability of the pair $(\mathbf{s}, \mathbf{s}_m)$ being real.

In sum, we train an separate encoders for $\mathcal{S}$, $\mathcal{S}_\text{M}$, $\mathcal{T}$, $\mathcal{T}_\text{M}$, separate decoders for $\mathcal{S}$, $\mathcal{S}_\text{M}$, $\mathcal{T}$, $\mathcal{T}_\text{M}$, and separate discriminator encoders for $\mathcal{D}_\mathcal{S}$, $\mathcal{D}_{\mathcal{S}_\text{M}}$, $\mathcal{D}_\mathcal{T}$, $\mathcal{D}_{\mathcal{T}_\text{M}}$. The system is joined cyclically such that all components involved in translations are trained simultaneously (Figure 2). Please see our supplemental material for details on the encoder and decoder architectures.

4.1. Loss Terms

GAN loss. We define our GAN loss following the comparative average ('relativistic') [16] variant of least squares GAN (LSGAN) [24], which has been shown to give stable results for image-to-image translation tasks [14, 15, 31, 37]:

$$\begin{aligned}\mathcal{L}_{GAN}(\Theta_D, \Theta_\mathcal{S}, \Theta_\mathcal{T}) = & (D_\mathcal{S}(d_\mathbf{s}, d_{\mathbf{s_m}}) - 1)^2 + \\ & (D_\mathcal{T}(d_\mathbf{t}, d_{\mathbf{t_m}}) - 1)^2 + \\ & D_\mathcal{S}(d_{\mathbf{t}'}, d_{\mathbf{t}'_m})^2 + D_\mathcal{T}(d_{\mathbf{s}'}, d_{\mathbf{s}'_m})^2,\end{aligned} \tag{1}$$

where Θ_D contains the parameters of $D_\mathcal{S}, D_\mathcal{T}, D_{\mathcal{S}_\text{M}}, D_{\mathcal{T}_\text{M}}$, and $\Theta_\mathcal{S}$ and $\Theta_\mathcal{T}$ contain parameters of the encoding and decoding parts in $\mathcal{S}$, $\mathcal{S}_\text{M}$ and $\mathcal{T}$, $\mathcal{T}_\text{M}$ respectively.

Our discriminators are also multi-scale, which has demonstrated better results on these tasks [10, 14, 20].

Cyclic loss. Unpaired image translation is an under-constrained problem even with $\mathcal{L}_{GAN}$. To further constrain it, we adopt the Zhu et al. [37] cyclic consistency which helps conserve characteristics of the input image (*e.g.*, the pose of the object). For this purpose, we could use a cycle energy loss on both the input images and their masks. This encourages an image mapped twice (from source to target to source) to be identical to the original image:

$$\mathcal{L}_{cyc}(\Theta_S, \Theta_T) = \|\mathbf{s} - \mathbf{s}''\|_1 + \|\mathbf{t} - \mathbf{t}''\|_1, \quad (2)$$

$$\mathcal{L}_{cyc_{\mathbf{m}}}(\Theta_S, \Theta_T) = \|\mathbf{s_m} - \mathbf{s}''_{\mathbf{m}}\|_1 + \|\mathbf{t_m} - \mathbf{t}''_{\mathbf{m}}\|_1, \quad (3)$$

where $\mathbf{s}'' = G_{\mathcal{T}}(h_{\mathbf{s}'}, h_{\mathbf{s}'_{\mathbf{m}}})$, with $h_{\mathbf{s}'} = E_{\mathcal{T}}(\mathbf{s}')$ and $h_{\mathbf{s}'_{\mathbf{m}}} = E_{\mathcal{T}_{\mathrm{M}}}(\mathbf{s}'_{\mathbf{m}})$. Likewise for $\mathbf{s}''_{\mathbf{m}}$.

While cyclic loss both aids the network in preventing mode collapse and in preserving attributes across domains, the term has limitations. The bijective constraint can hinder shape change [10, 37] and texture transfer. Small irrelevant details must be preserved to perfectly reconstruct the original image. Instead, we focus on reconstructing a much simpler task: the original segmentation. In our results, we will show empirically that minimizing only $\mathcal{L}_{cyc_{\mathbf{m}}}$ is enough to achieve the goal initially desired by cycle energies.

We find that eliminating the RGB reconstruction loss improves the quality of generated images (Table 3). The remaining loss terms provide adequate stability to the network to prevent mode collapse. Furthermore, unlike previous methods which do not require cyclic reconstruction [3], our method can handle shape change.

Identity loss. Furthermore, our approach is made more robust via enforcing an identity loss which encourages the generated images to be identical to the inputs in the case where source and target domains are the same:

$$\begin{aligned}\mathcal{L}_{id}(\Theta_S, \Theta_T) =& \|\mathbf{s} - G_{\mathcal{T}}(h_{\mathbf{s}}, h_{\mathbf{s_m}})\|_1 + \\ & \|\mathbf{t} - G_{\mathcal{S}}(h_{\mathbf{t}}, h_{\mathbf{t_m}})\|_1, \quad (4)\\ \mathcal{L}_{id_{\mathbf{m}}}(\Theta_S, \Theta_T) =& \|\mathbf{s_m} - G_{\mathcal{T}_{\mathrm{M}}}(h_{\mathbf{s}}, h_{\mathbf{s_m}})\|_1 + \\ & \|\mathbf{t_m} - G_{\mathcal{S}_{\mathrm{M}}}(h_{\mathbf{t}}, h_{\mathbf{t_m}})\|_1. \quad (5)\end{aligned}$$

Context loss. Jointly optimizing the previous three losses enables shape change for image-to-image translation; however, there is no mechanism encouraging only foreground objects of interest to be altered, *e.g.*, in the giraffe to sheep problem, one should be able to change a giraffe to the corresponding sheep without altering the background. One way to achieve such property is to enforce the initial background pixels in $\mathbf{s}$ and given by $\mathbf{s_m}$ not to change in $\mathbf{t}'$. However, doing so would be in opposition with our initial motivation of performing significant shape changes across domains as, in that scenario, pixels which were initially in the background of $\mathbf{s}$ could end up being in the foreground of $\mathbf{t}'$. For this reason, we instead enforce the *intersection* of backgrounds in $\mathbf{s_m}$ and $\mathbf{t}'_{\mathbf{m}}$ not to change, which allows for shape change while encouraging the background to be conserved as in InstaGAN

$$\begin{aligned}\mathcal{L}_{ctx}(\Theta_S, \Theta_T) =& \|(\mathbf{s} - \mathbf{t}') \odot w(\mathbf{s_m}, \mathbf{t}'_{\mathbf{m}})\|_1 + \\ & \|(\mathbf{t} - \mathbf{s}') \odot w(\mathbf{t_m}, \mathbf{s}'_{m})\|_1, \quad (6)\end{aligned}$$

where $\odot$ represents Hadamard product and $w(\mathbf{s_m}, \mathbf{t}'_{\mathbf{m}}) = 1 - \max(\tilde{\mathbf{s}}_{\mathbf{m}}, \tilde{\mathbf{t}}'_{\mathbf{m}})$ with $\tilde{\mathbf{s}}_{\mathbf{m}}$ being the binary, rounded version of the initial mask. Furthermore, since this constraint operates as a soft constraint, the network may make small changes outside the segmentation that enhance the plausibility of the image. For instance, filling gaps in between segmentation or slightly changing the lighting tone or vegetation in an image.

Feature matching loss. Finally, we define a feature matching loss which we find to encourage sharper and more realistic images. This loss encourages feature representations computed from $\{\mathbf{s}'_i\}_{i=1}^{N_t}$ to be close in terms of the $L2$ norm with respect to the features computed from $\{\mathbf{s}_i\}_{i=1}^{N_s}$, where N_s and N_t are the number of datapoints for domains $\mathcal{S}$ and $\mathcal{T}$ respectively. Such feature representations are extracted in this case from $D_{\mathcal{S}}$.

This loss has been shown to help produce good generated images [28, 31, 10, 30]. The intuitive reason why this works is that we wish neurons in the source domain discriminator to be activated in the same way between generated source images and real ones. This ends up reinforcing the realism of the generator, $G_{\mathcal{S}}$ in this case.

Formally, we express this via the following loss:

$$\begin{aligned}\mathcal{L}_{FM}(\Theta_S, \Theta_T) = \sum_{l=1}^{L-1} & \mathbb{E}_{\mathbf{s}\sim B}(\|D^l_{\mathcal{S}}(\mathbf{s}') - D^l_{\mathcal{S}}(\mathbf{s})\|_2^2) + \\ & \mathbb{E}_{\mathbf{t}\sim B}(\|D^l_{\mathcal{T}}(\mathbf{t}') - D^l_{\mathcal{T}}(\mathbf{t})\|_2^2), \quad (7)\end{aligned}$$

where L is the number of layers in the discriminator architecture, B is the mini-batch, and D^l is the feature activation of the corresponding discriminator at layer l. Moreover, since we use a small batch size, directly enforcing Eq. 7 may be less valid for a given pair of s'_i and s_i. We introduce the novel contribution of maintaining an exponential moving average for feature representations of the real images and use them instead in Eq. 7.

4.2. Improvements over Previous Works

Our work is inspired by recent works such as ContrastGAN [21] and InstaGAN [2] which augment CycleGAN with a segmentation map. We make several important changes, each of which meaningfully improves the translation quality as we show in the ablation study (Sec. 5.1). For clarity, we list all changes below.

Additions. Over losses in InstaGAN, we add relativistic average discriminators [16] via least squares GAN [24], multi-scale discriminators, and feature matching loss. We also maintain an exponential moving average over the feature match loss to learn the feature statistics of real images across the entire dataset despite the small minibatch size.

Mask supervision	Giraffe	Sheep	Bear	Elephant	Fire Hydrant	Toilet
No supervision	0.000	0.000	0.000	0.000	0.000	0.000
Bounding box (GT)	0.381	0.612	0.643	0.638	0.602	0.644
Cosegmentation	0.467	0.570	0.663	0.616	0.077	0.032
DeepGrabCut+BBox	0.647	0.665	0.824	0.768	0.727	0.726
Full	1.000	1.000	1.000	1.000	1.000	1.000

Table 1: To evaluate the accuracy of various segmentation methods we compute the IOU across several categories.

Subtractions. We remove the '*insta*nce' from InstaGAN: We only provide one mask per image which contains the intersection of all per-instance masks, instead of one mask per instance in the image framed as recursive segmentation. This reduces complexity and avoids the need for expensive explicit instance segmentation. We show that this still achieves high quality results. Further, we do not enforce a cyclic loss on the RGB reconstruction and only on the mask reconstruction, which again saves computation.

5. Experiments

Datasets and supervision. We use the Microsoft CoCo dataset [22] as it includes human-labeled binary segmentation masks for objects. We select pairs of object classes to translate between (number of training images): giraffe (2647), sheep (1594), bear (1009), elephant (2232), airplane (3083), bird (3362), toilet (3502), fire hydrant (1797).

From the human-labeled masks, we derive bounding boxes as an imprecise form of supervision. Then, given the bounding box, we also generate an intermediate-precision supervision. There are several papers that attempt to generate segmentations of objects given bounding boxes, such as Grabcut [26], DeepGrabcut [33] and Simply Does It [17]. We perform Deep Grabcut [33] inside the bounding box to improve the supervision.

Finally, given just class label information (which is implicit in our choices of S and T, we can derive segmentations using modern cosegmentation techniques as a form of imprecise supervision. We use the framework from Chen et al. [5] pretrained on Pascal VOC 2012 [9] to derive segmentations from the MS CoCo dataset. We randomly select pairs of images from the same class. If an image has an empty mask through any of these supervision generation processes, then we reject it from training.

Table 1 shows IOU scores with respect to manual image segmentation labels for some of the classes used in our experiments. We can see high amounts of inter-class variability, *e.g.*, giraffes are non-rectangular, and so are poorly approximated by bounding boxes.

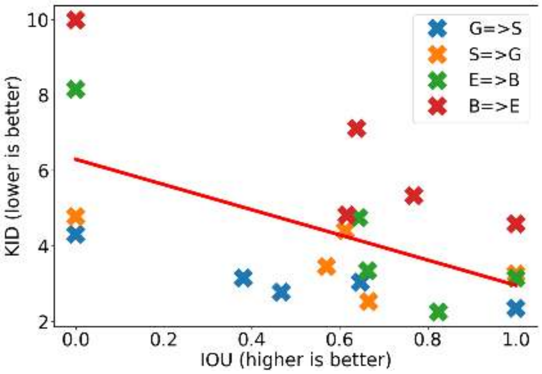


Figure 3: Correlation between IOU and KID scores. IOU is computed with respect to manually labeled segments, and is a rough indicator of segmentation quality. We can see here that there is a modest trend towards improved KID (more realistic images), as segmentation quality improves.

Metric. We use the recently-proposed Kernel Inception Distance (KID) [4] as a measure of image quality, as it has been shown to outperform the previous most common used metric, FID [12], especially for smaller sets. KID works by computing the squared Maximum Mean Discrepancy (MMD) between the feature representations of two sets of images, in our case real and generated images from the target domain. Feature representations for each set are extracted from the last hidden layer of the inception architecture [29] pretrained on ImageNet [8].

5.1. Results

We show results for differing levels of supervision on image-to-image translation in Table 2. The top grouping of methods are scores for the full-supervision case where segmentation masks are provided. The second group contains scores for less precise supervised approaches using only bounding box annotations and automatically inferred masks. The third group contains models where ground-truth segmentation masks from a different dataset (PASCAL VOC 2012 [9]) have been used to pre-train a cosegmentation network [5] which then provides us with training masks for our experiments. Finally, the last part contains models where no supervision is given (*i.e.*, only RGB images from different domains are given as input).

Overall fully supervised models using feature matching are the most competitive. We see a trend in the correlation between KID and IOU scores (Figure 3). The type of mask used to be chosen based on desired result quality and dataset cost requirements. Fully unsupervised approaches such as CycleGAN perform worst on average.

Ablation study. We evaluate the role of each improvement to our InstaGAN-motivated architecture (Fig. 4). We find that augmenting with a relativistic discriminator helps

Table 2: Kernel Inception Distance×100±std×100. for different supervision levels (top to bottom is most to least). Lower is better. Abbreviations: (G)iraffe, (S)heep, (B)ear, (E)lephant, (F)ire hydrant, (T)oilet.

Model	Object Mask Supervision	$G \to S$	$S \to G$	$B \to E$	$E \to B$	$F \to T$	$T \to F$
Proposed	Human per-pixel	**2.95 ± 0.34**	3.07 ± 0.28	2.92 ± 0.09	4.58 ± 0.28	**5.07 ± 0.43**	5.17 ± 0.17
Proposed	Cosegmentation shape prior	2.86 ± 0.21	3.55 ± 0.13	3.32 ± 0.15	4.47 ± 0.39	10.66 ± 0.61	6.63 ± 0.29
Proposed	DeepGrabCut from human bounding box	3.00 ± 0.15	**3.03 ± 0.14**	**2.52 ± 0.09**	**4.45 ± 0.41**	6.02 ± 0.46	**5.05 ± 0.19**
Proposed	GrabCut from human bounding box	3.56 ± 0.20	3.77 ± 0.10	3.10 ± 0.28	4.77 ± 0.26	7.85 ± 0.57	5.82 ± 0.28
Proposed	Human bounding box	3.13 ± 0.21	4.73 ± 0.14	3.37 ± 0.27	6.62 ± 0.28	6.46 ± 0.46	5.93 ± 0.16
CycleGAN	None	4.39 ± 0.29	4.60 ± 0.33	7.54 ± 0.58	10.33 ± 0.52	10.39 ± 0.63	7.03 ± 0.25

Table 3: Kernel Inception Distance×100±std×100 for the different ablation studies conducted. Lower is better. Abbreviations: (G)iraffe, (S)heep.

Model	$G \to S$	$S \to G$
InstaGAN	3.32 ± 0.15	3.08 ± 0.30
+ Relativistic Discriminator	3.15 ± 0.16	3.18 ± 0.11
+ Multiscale Discriminator	3.22 ± 0.64	**3.00 ± 0.30**
− RGB Cyclic Loss	3.02 ± 0.20	3.35 ± 0.13
+ Feature Match Loss	3.30 ± 0.51	3.25 ± 0.13
+ Feature Match Loss with EMA average	**2.95 ± 0.34**	3.07 ± 0.28

improve the shape change of generated images *e.g.*, generated giraffes look more realistic (3rd column in the right part of Fig. 4). We further find that removing the cycle energy on the RGB images produces comparable results, which indicates that the weaker cycle energy form which acts only on the masks is enough to conserve characteristic of the original image. Finally, we find that adding multiscale discriminators and an exponentially averaged feature matching loss further improves the performance. Our qualitative findings are supported quantitatively by KID scores (Table 3).

6. Discussion

We posit that ground truth pixel-wise mask segmentations provide two critical pieces of information to the network: localization of the object(s) in the image, and disentanglement of the (textureless) shape characteristics of the object. Bounding boxes provide the location of objects but do not reveal any information about their shape. Rectangles possess poor IOU with the ground truth segmentations of animals such as giraffes (Table 1).

The lower result quality from bounding boxes demonstrate that the shape hints provided by richer segmentation are important. However, lower quality masks generated from Deep GrabCut on bounding boxes [33] still produce good results both quantitatively and qualitatively (Fig. 6).

Full supervision cyclic failure. In our supplemental, we demonstrate the bird to airplane dataset. In this case, Deep GrabCut outperforms the fully supervised masks; even CycleGAN performs well. The human-provided segmentations hurt network performance both qualitatively and in terms of KID score. We posit that this is because of images containing flocks of birds in the distance, which produce many small clustered mask regions which are hard to translate or reproduce cyclically. For these regions, that imprecise supervision is less specific is beneficial.

Object complexities. Small objects in general can be difficult to segment effectively (both for human labelers and for automatic or semi-automatic methods). Localizing small and distant objects such as kitchenware and utensils in an unsupervised manner is difficult and requires at least some mask supervision. Furthermore, small objects that occur with common objects, such as bottles with humans, can also be unrecognized by cosegmentation approaches.

Additionally, approaches can struggle even in the supervised case when two classes appear in the same image, *e.g.*, a bus-to-car translation where images of buses often also contain a car. Therefore, the generator can learn the trivial solution which switches car and bus segmentation labels; this satisfies the bijective requirement by exploiting the RGB input to re-segment the object.

Backgrounds. Controlling background differences remains a challenging problem: the context loss constraint on background changes must be weighted with other losses, which allows the system some freedom. This can be desirable when modifying the background yields a more believable image, such as in changing content close to the mask to better cope with object shape changes. These cases tend to rely on the backgrounds already having at least some similarity, e.g., in animal conversions with natural landscape backgrounds. However, it can also yield poor results when backgrounds differ more. Our fire hydrant/toilet examples show generally higher KID scores as backgrounds from street scenes merge into bathrooms (Fig. 6, bottom). Increasing the context loss significantly can slow down or halt training as the discriminator becomes increasingly reliant on the background patches which the generator is unable to change. Future work may seek to find a better soft

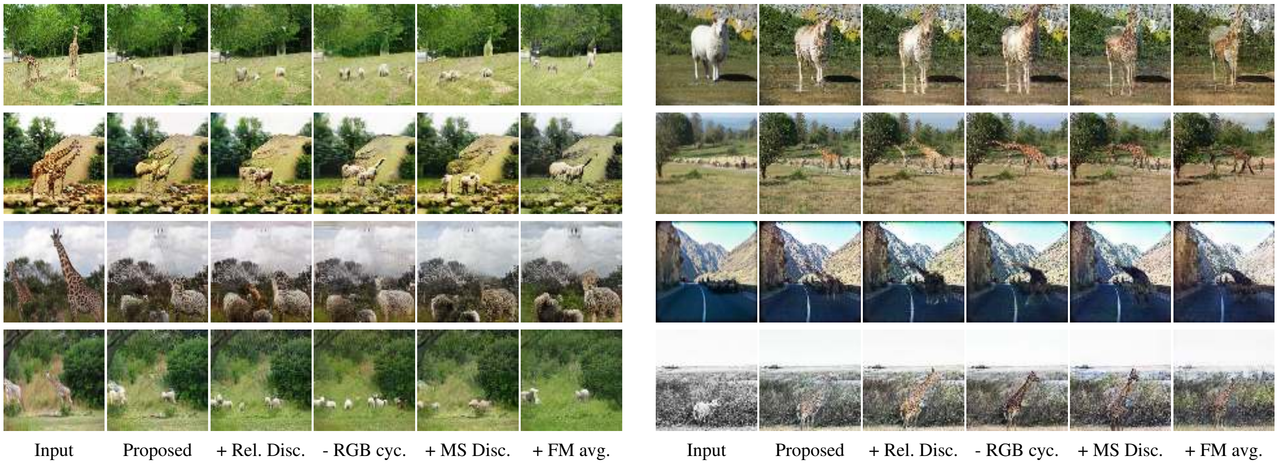


Figure 4: Qualitative results for an ablation of our losses, where we add a relativistic discriminator (+ Rel. Disc.), multiscale discriminator (+ MS Disc.), feature matching EMA average (+ FM avg.), and subtract an RGB cyclic loss (- RGB cyc.). Left: Giraffe → Sheep. Right: Sheep → Giraffe.

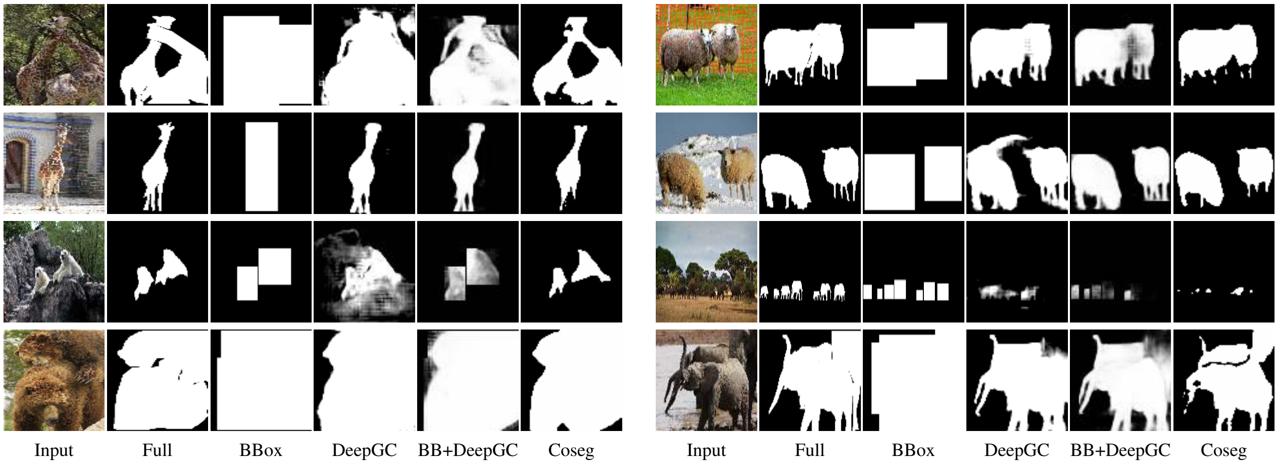


Figure 5: Examples for different supervision and weak supervisions used for the class Giraffe, Sheep, Bear, and Elephant.

attention mechanism than concatenation.

6.1. Additional experiments

GrabCut. We experimented with GrabCut [26] vs. Deep GrabCut, which presents a classic speed/performance tradeoff. While less computationally expensive, GrabCut produced worse KID score in almost all cases compared to Deep GrabCut [33], except for 'elephant to bear'.

Explicit segmentation discriminator. We also tried adding an additional discriminator which only looked at the segmentation channel. While it did improve mask quality, it did so at the expense of RGB quality. Furthermore, it performed worse with imprecise supervised masks and had a habit of adding additional instances of the animal.

CycleGAN with mask concatenated. Concatenating the segmentation masks to the RGB channels in CycleGAN led the network to produce all-black masks. By focusing only on the RGB channels, CycleGAN minimized the L1 cyclic loss. Since most of the background mask region is black, a trivial minimizer makes the entire mask image black.

Unsupervised attention. Using automatically-learned attention maps [25, 6] turned out not to be robust to shape change. These algorithms constrain the generator to only modify attended regions in the image; however, if the source and target domains are not of the same shape, such an assumption can be constraining. For example, it is challenging to transform a sheep to a giraffe by only modifying pixels covered by the sheep. Consequently, these networks end up attending to the entire image to converge, which is simi-

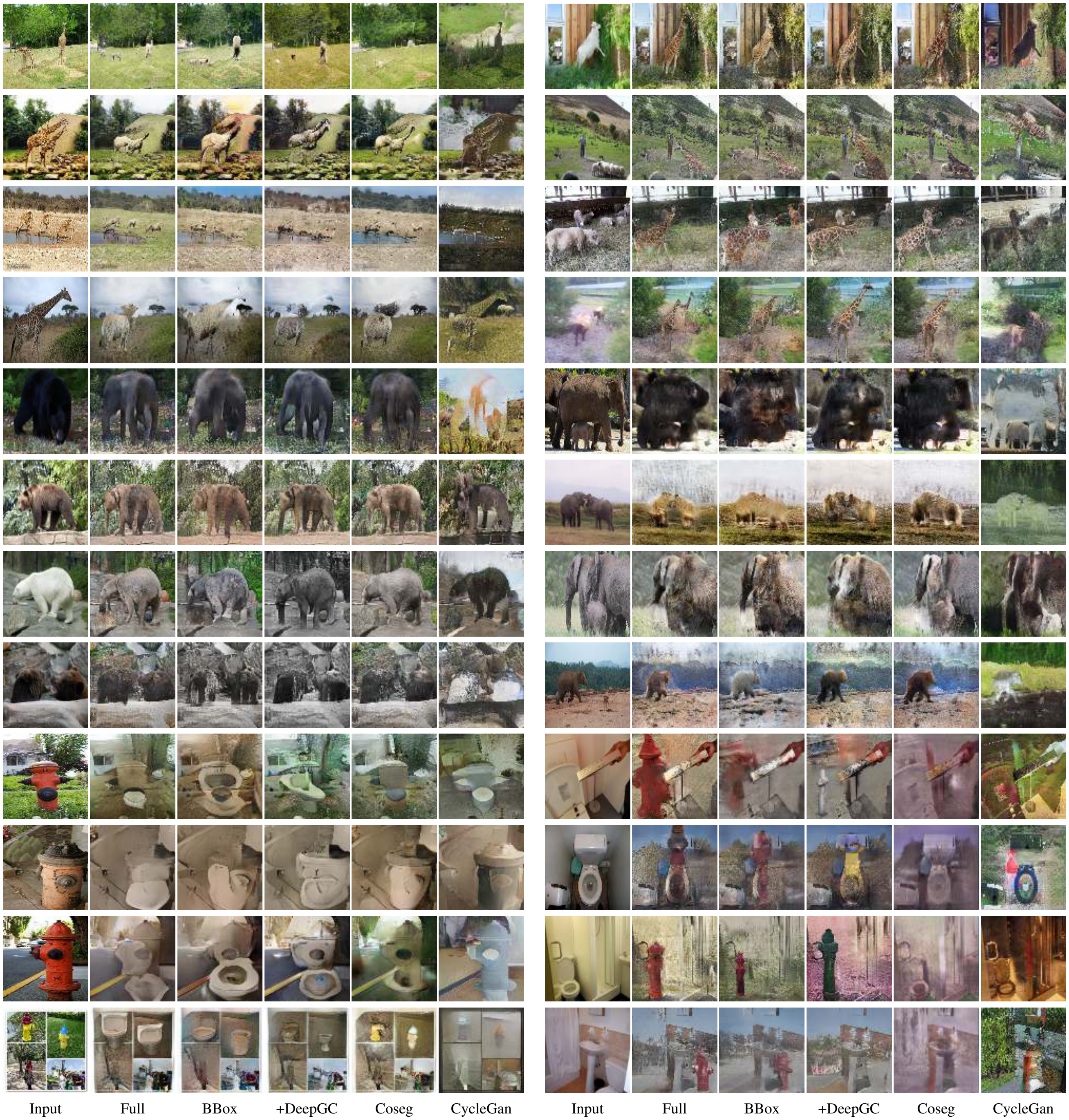


Figure 6: Influence of different levels of supervision on image-to-image translation under significant shape change.

lar to the CycleGAN setting.

7. Future Work

While we have shown that imprecise supervision can be helpful for challenging image-to-image translation tasks, more work is needed to enable robust shape change in a fully unsupervised manner (Fig. 6; CycleGAN struggles). Recent work has made strides enabling unsupervised shape change [10]; however, they still alter the background significantly. Unsupervised attention-based approaches successfully conserve the background [25], but can only deal with limited shape change. Future work should combine these properties to enable instance-aware image translation while simultaneously disentangling the background.

Moreover, our approach is agnostic to the precision, type, and creation method of the segmentation masks. Future work may achieve higher performance by more closely integrating these differences into the model. We could also train on several classes at once as per other recent works [1, 7, 14, 19, 21, 27].

References

[1] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018. 9

[2] Anonymous. Instance-aware image-to-image translation. In *Submitted to International Conference on Learning Representations (under review)*, 2019. under review. 1, 2, 3, 4

[3] S. Benaim and L. Wolf. One-sided unsupervised domain mapping. In *NIPS*, pages 752–762, 2017. 4

[4] M. Bińkowski, D. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *ICLR*, 2018. 5

[5] H. Chen, Y. Huang, and H. Nakayama. Semantic aware attention based deep object co-segmentation. In *ACCV*, 2018. 1, 2, 5

[6] X. Chen, C. Xu, X. Yang, and D. Tao. Attention-GAN for object transfiguration in wild images. In *ECCV*, 2018. 2, 7

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Computer Vision and Pattern Recognition*, 2018. 9

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 5

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 5

[10] A. Gokaslan, V. Ramanujan, D. Ritchie, K. I. Kim, and J. Tompkin. Improved shape deformation in unsupervised image to image translation. In *ECCV*, 2018. 2, 3, 4, 8

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Klambauer. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NIPS*, 2017. 5

[13] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang. Co-attention CNNs for unsupervised object co-segmentation. *International Joint Conference on Artificial Intelligence*, 2018. 2

[14] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3, 9

[15] P. Isola, J. Zhu, T. Zhou, and A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[16] A. Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*, July 2018. 3, 4

[17] A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, and B. Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017. 5

[18] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *JMLR*, 2017. 2

[19] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. *arXiv preprint arXiv:1808.00948*, 2018. 9

[20] M. Li, H. Huang, L. Ma, W. Liu, T. Zhang, and Y.-G. Jiang. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. *arXiv preprint arXiv:1807.08536*, abs/1807.08536, 2018. 3

[21] X. Liang, H. Zhang, and E. P. Xing. Generative semantic manipulation with contrasting GAN. *arXiv preprint arXiv:1708.00315*, abs/1708.00315, 2017. 1, 2, 4, 9

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft CoCo: Common objects in context. In *ECCV*. Springer, 2014. 5

[23] S. Ma, J. Fu, C. W. Chen, and T. Mei. DA-GAN: Instance-level image translation by deep attention generative adversarial networks. In *CVPR*, 2018. 2

[24] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *ICCV*, 2017. 3, 4

[25] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim. Unsupervised attention-guided image to image translation. *arXiv preprint arXiv:1806.02311*, 2018. 2, 7, 8

[26] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, 2004. 5, 7

[27] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Moressi, F. Cole, and K. Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. *arXiv preprint arXiv:1711.05139*, 2017. 9

[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *NIPS*, 2016. 4

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016. 5

[30] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NIPS*, 2018. 4

[31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 3, 4

[32] X. Xia and B. Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017. 2

[33] N. Xu, B. L. Price, S. Cohen, J. Yang, and T. S. Huang. Deep grabcut for object selection. *arXiv preprint arXiv:1603.04042*, abs/1707.00243, 2017. 1, 2, 5, 6, 7

[34] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C. J. Kuo. Show, attend and translate: Unsupervised image translation with self-regularization and attention. *arXiv preprint arXiv:1806.06195*, 2018. 2

[35] J. Yang, A. Kannan, D. Batra, and D. Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *arXiv preprint arXiv:1703.01560*, 2017. 2

[36] Z. Ye, F. Lyu, J. Ren, Y. Sun, Q. Fu, and F. Hu. DAU-GAN: Unsupervised object transfiguration via deep attention unit. In *Advances in Brain Inspired Cognitive Systems*, 2018. 2

[37] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2, 3, 4

Exploring the Spectrum of Mask Supervision for Unpaired Image-to-Image Translation Supplementary Material

Aaron Gokaslan

Brown University

Table 1: Encoding architecture for the generator. ‘Conv.’ is convolutional layer; ‘Res.’ is residual block; ‘InstNorm’ is instance normalization; ‘Act.’ is activation function.

Layer	#Filters	Size	Stride	InstNorm	Act.
Conv.	64	7×7	1	✓	ReLU
Conv.	128	3×3	2	✓	ReLU
Conv.	256	3×3	2	✓	ReLU
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident
Res.	256	3×3	1	✓	Ident

1. Network Architectures

Our networks are constructed from similar components to Cycle GAN [3]. The encoder consists of three convolutional downsampling layers and nine residual blocks (Table 1). The decoding architecture of the generator is given by Table 3.

The discriminator also consists of a similar encoder-decoder structure. The encoding architecture is in this case give in Table 4, while the decoder is given in Table 5.

2. Systems

All experiments were run using either two NVidia 1080 GPUs or a single NVidia 1080 Ti GPU, written in Tensorflow [1] using the Tensorpack [2] library. We will release our code and data as open source.

Table 2: Residual block architecture.

Layer	#Filters	Size	Stride	InstNorm	Act.
Conv.	256	3×3	1	✓	ReLU
Conv.	256	3×3	1	✓	Ident

Table 3: Decoding architecture for the generator. ‘Deconv.’ refers to a transposed convolution. Blue row applies when decoding RGB, Orange row applies when decoding a mask.

Layer	#Filters	Size	Stride	InstNorm	Act.
Deconv.	128	3×3	2	✓	ReLU
Deconv.	64	3×3	2	✓	ReLU
Conv.	3	7×7	2	-	Tanh
Conv.	1	7×7	2	-	Tanh

Table 4: Encoding architecture for the Discriminators. ‘LReLU’ denotes Leaky ReLU with a factor of 0.2.

Layer	#Filters	Size	Stride	InstNorm	Act.
Conv.	64	4×4	2	-	LReLU
Conv.	128	4×4	2	✓	ReLU
Conv.	256	4×4	2	✓	ReLU

Table 5: Decoding architecture for the Discriminators.

Layer	#Filters	Size	Stride	InstNorm	Act.
Conv.	512	4×4	1	✓	ReLU
Conv.	1	4×4	1	-	Ident

3. Additional results

We show qualitative results on the Airplane and Bird dataset for their bidirectional translations (Figure 1) as well as Horse to Zebra dataset (Figure 2); plus additional qualitative results for the datasets used in the main paper (Giraffe and Sheep—Figure 3; Bear and Elephant—Figure 4; Toilet and Fire Hydrant—Figure 5). All images shown are drawn from the test dataset; we sample images at different object scales and with different object configurations to provide a representation of overall result quality. We also reject images with close-ups, e.g., where we see only the wing of a plane or only the face of a sheep.

For the special case of Fire Hydrant to Toilet (Figure 5), we notice that the background endures severe changes, moreover the generated foreground does not generally blend correctly with the background. This is because these specific classes (Fire Hydrant and Toilet) are generally present in completely different scenes: the first one is on the street while the second one is in a restroom. Such environment discrepancy obliges the generator to heavily alter the background in order to convince the corresponding discriminator.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 1

[2] Y. Wu et al. Tensorpack. https://github.com/tensorpack/, 2016. 1

[3] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1

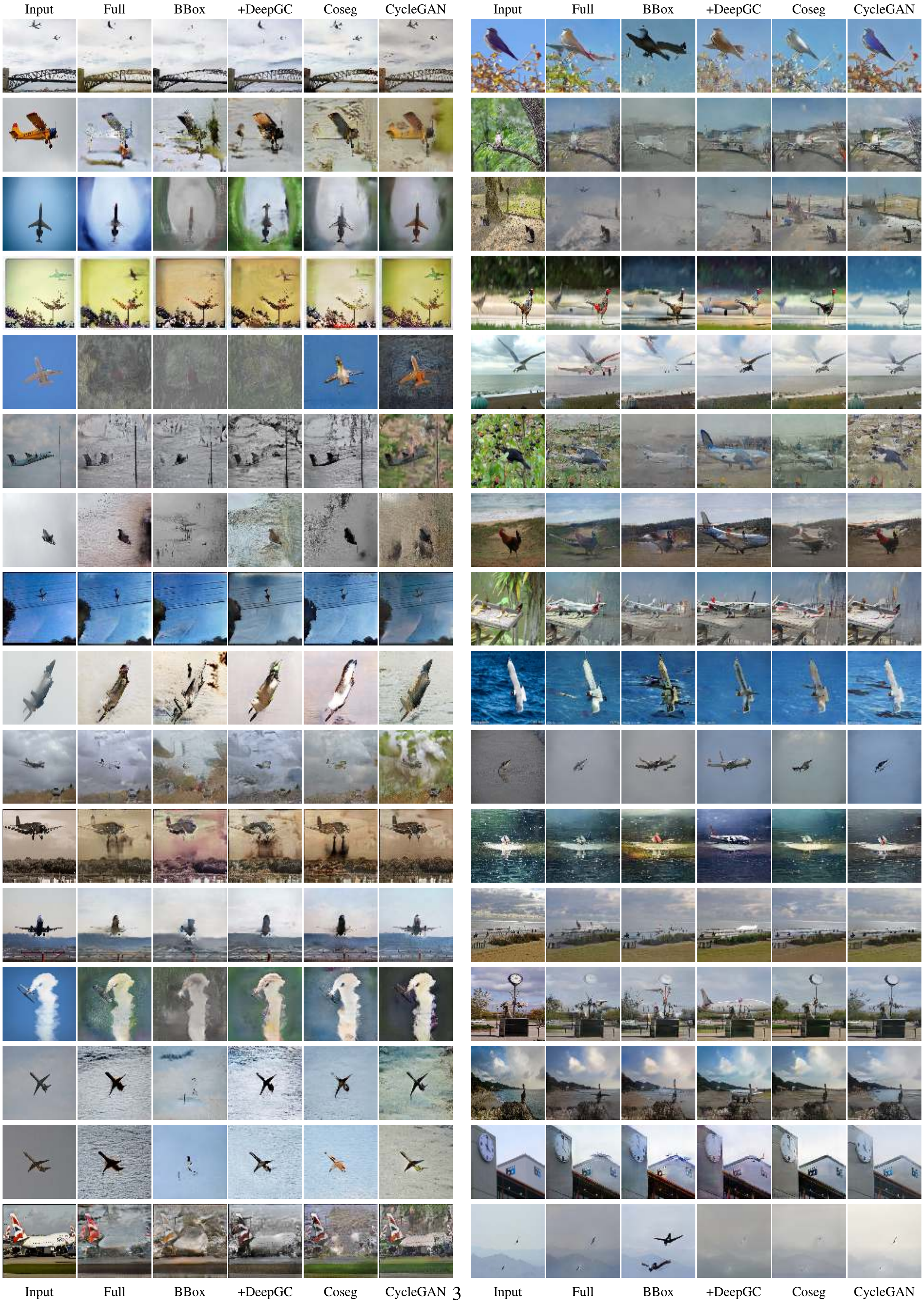


Figure 1: Additional results. (Left): Airplane → Bird, (Right): Bird → Airplane.

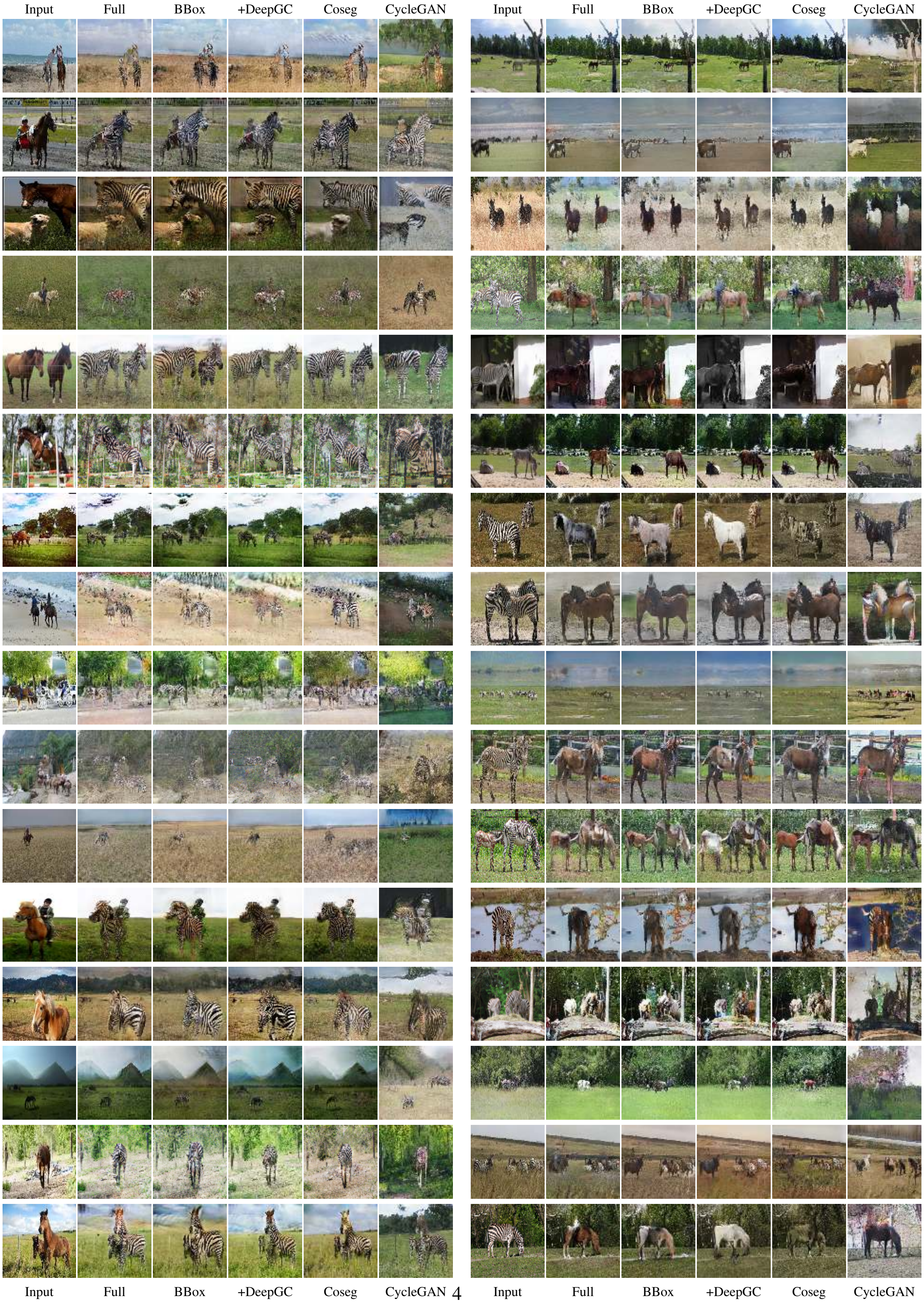


Figure 2: Additional results. (Left): Horse → Zebra, (Right): Zebra → Horse.

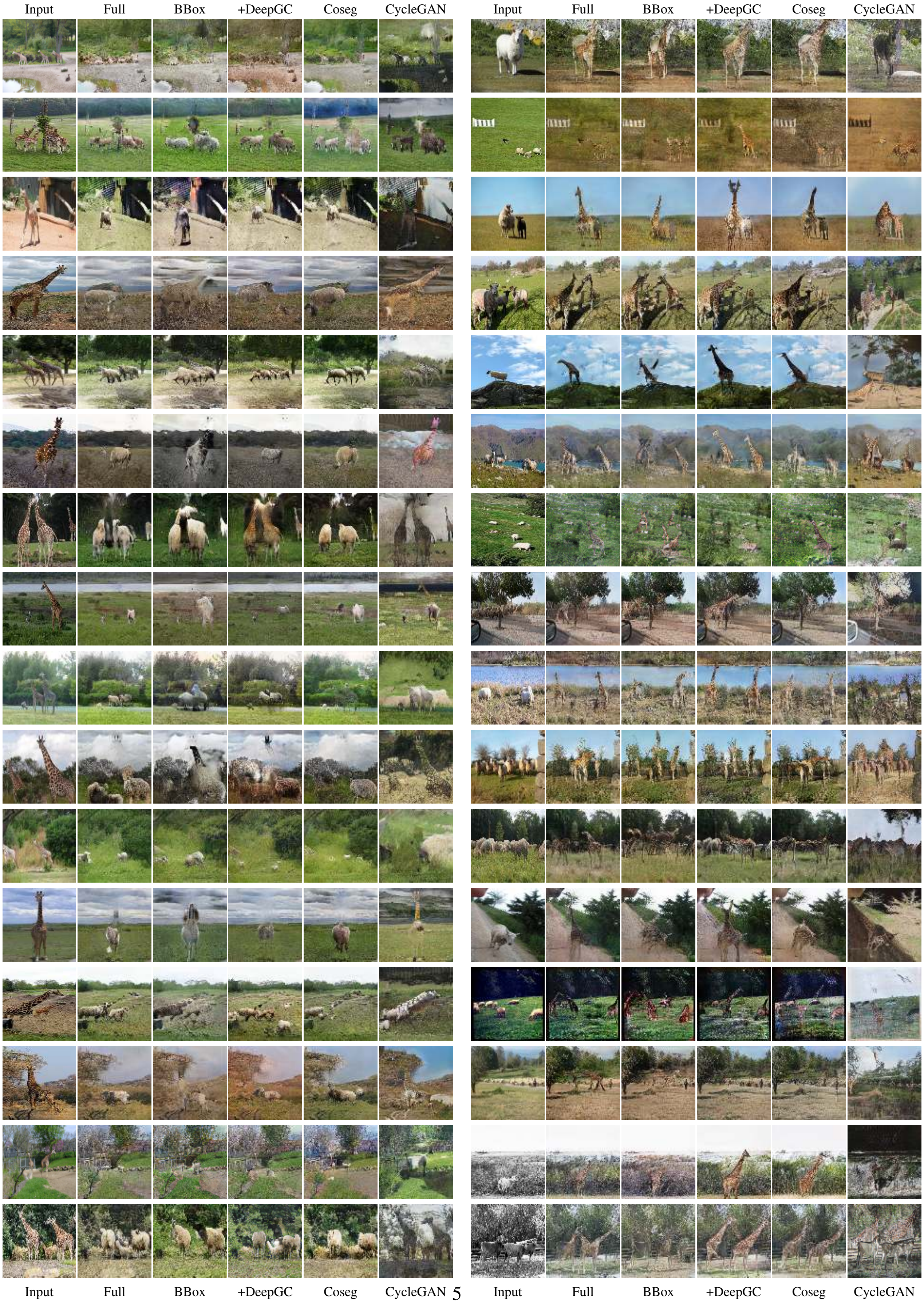


Figure 3: Additional results. (Left): Giraffe → Sheep, (Right): Sheep → Giraffe.

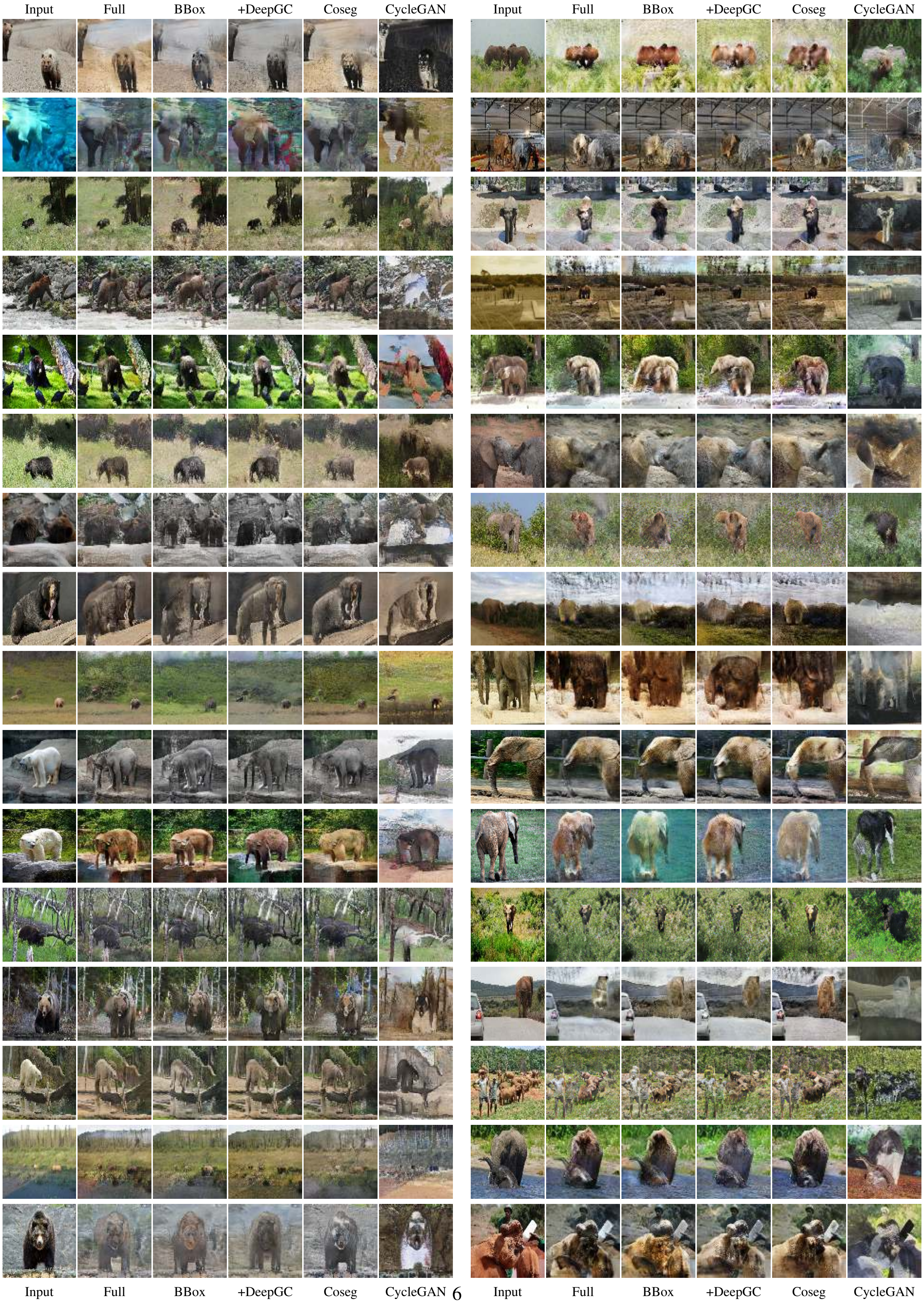


Figure 4: Additional results. (Left): Bear → Elephant, (Right): Elephant → Bear.

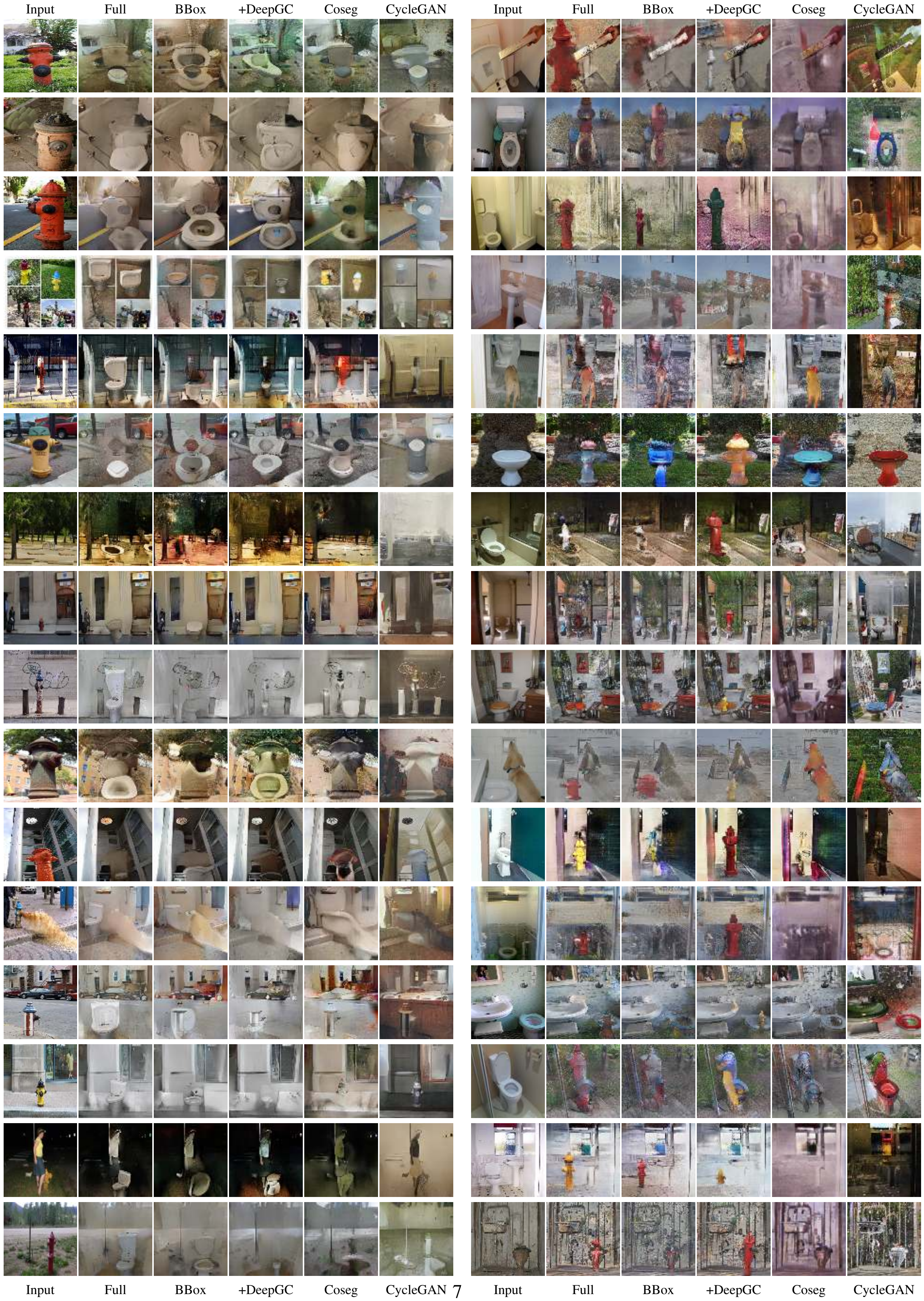


Figure 5: Additional results. (Left): Fire hydrant → Toilet, (Right): Toilet → Fire hydrant.