BalaGAN: Image Translation Between Imbalanced Domains via Cross-Modal Transfer

Or Patashnik[1] **Dov Danon**[1] **Hao Zhang**[2] **Daniel Cohen-Or**[1]

[1]Tel-Aviv University, Israel [2]Simon Fraser University, Canada

Abstract

State-of-the-art image-to-image translation methods tend to struggle in an *imbalanced* domain setting, where one image domain lacks richness and diversity. We introduce a new *unsupervised* translation network, *BalaGAN*, specifically designed to tackle the domain imbalance problem. We leverage the *latent modalities* of the richer domain to turn the image-to-image translation problem, between two imbalanced domains, into a *balanced, multi-class*, and *conditional* translation problem, more resembling the style transfer setting. Specifically, we analyze the source domain and learn a decomposition of it into a set of latent modes or classes, without any supervision. This leaves us with a multitude of *balanced cross-domain* translation tasks, between all pairs of classes, including the target domain. During inference, the trained network takes as input a source image, as well as a reference or style image from one of the modes as a condition, and produces an image which resembles the source on the pixel-wise level, but shares the same mode as the reference. We show that employing modalities within the dataset improves the quality of the translated images, and that BalaGAN outperforms strong baselines of both unconditioned and style-transfer-based image-to-image translation methods, in terms of image quality and diversity.

1 Introduction

Image-to-image translation is a central problem in computer vision and has a wide variety of applications including image editing, style transfer, data enrichment, image colorization, among others. Acquiring labeled pairs of source and target domain images is often hard or impossible, thus motivating the development of unsupervised methods (Zhu et al., 2017; Huang et al., 2018; Kim et al., 2020; Park et al., 2020; Lira et al., 2020; Liu et al., 2019; Choi et al., 2020). However, these methods are often lacking in quality or robustness to domain variations. Indeed, in most unsupervised approaches, there is an implicit assumption of "approximate symmetry" between the translated domains, in term of data quantity or variety. With this assumption, the source and target domains are treated each as *one-piece*, without fully leveraging the variety within either of them. In reality, most datasets are imbalanced across different categories, e.g., ImageNet (Deng et al., 2009) contains many more images of dogs than of wolves. As image-to-image translation can be used to enrich some domains by utilizing others, improving these methods, in the imbalanced setting in particular, can play a critical role in resolving the ubiquitous "data shortage" problem in deep learning.

In this paper, we present *BalaGAN*, an *unsupervised* image-to-image translation network specifically designed to tackle the domain imbalance problem where the source domain is much richer, in quantity and variety, than the target one. Since the richer domain is, in many cases, *multi-modal*, we can leverage its *latent* modalities. To do this, we turn the image-to-image translation problem, between two imbalanced domains, into a *balanced, multi-class,* and *conditional* translation problem, akin to style transfer. Our key observation is that the performance of a domain translation network can be significantly boosted by (i) disentangling the complexity of the data, as reflected by the natural modalities in the data, and (ii) training it to carry out a multitude of varied translation tasks instead of a single one. BalaGAN fulfills both criteria by learning translations between *all pairs* of source domain modalities *and* the target domain, rather than only between the full source and target domains. This way, we are taking a *balanced* view of the two otherwise imbalanced domains.

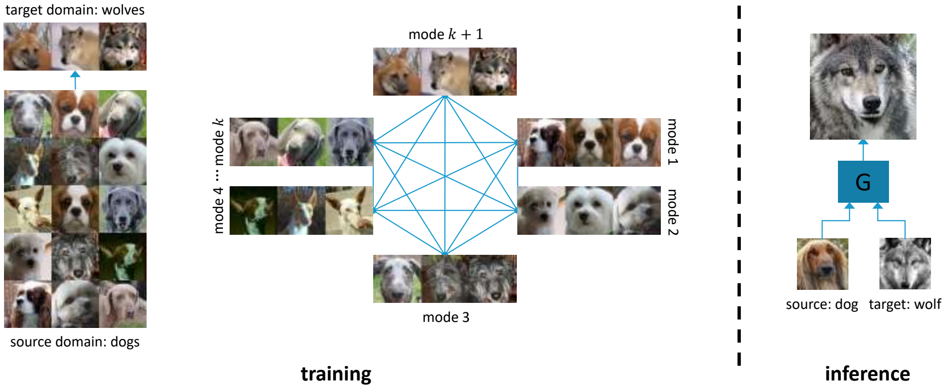


Figure 1: Our image translation network, BalaGAN, is designed to handle imbalanced input domains, e.g., a set of dog images that is much richer than that of wolves. We “balance” the domains by converting them into multiple modes reflecting their “styles” and train a GAN over *all mode pairs* to learn a multitude of intra- and inter-mode cross-translations. During inference, the network takes a source (e.g., a dog) and a reference image (e.g., a wolf) to produce a new image following the “style/mode” of the reference while resembling the source in a pixel-wise manner.

More importantly, enforcing the network to learn such a richer set of translations leads to improved results, and in particular, a better and more diverse translation to the target domain.

Specifically, let us assume that the source domain A, which is significantly richer than the target domain B, consists of multiple mode classes. We train a *single* GAN translator G with respect to all pairs of modes (see Figure 1). During inference, the trained network takes as input a source image x, as well as a *reference image* y from one of the modes as a condition, and produces an image $G(x, y)$. This image resembles x on the *pixel-wise* level, but shares the same mode (or style) as y. To realize our approach, we develop means to find the latent data modalities without any supervision and a powerful generator for the task of conditional, multi-class image-to-image translation. Our translator is trained adversarially with two discriminators, each aiming to classify a given image to its corresponding mode, with one trained on real images only. The generator is trained to produce meaningful content and style representations, and combine them through an AdaIN layer. While this architecture bears resemblance to multi-class translation networks such as FUNIT (Liu et al., 2019) and StarGAN (Choi et al., 2020), it should be emphasized that unlike these methods, we learn the latent modalities, and use *transductive learning*, where the target domain participates in the training.

We show that reducing the imbalanced image translation problem into a cross-modal one achieves comparable or better results compared to any unsupervised translation method we have tested, including the best performing and most established ones, since they do not exploit the latent modalities within the source domain. We analyze the impact of the extracted latent modalities, perform ablation studies, and extensive quantitative and qualitative evaluations, which are further validated through a perceptual user study. We further show the potential of our cross-modal approach for boosting the performance of translation in balanced setting.

2 Related Work

Modern unsupervised image-to-image translation methods use GANs (Goodfellow et al., 2014) to generate plausible images in the target domain, conditioned on images from a source domain. Such methods are unsupervised in the sense that no pairs between the source and target domain are given. Some works (Zhu et al., 2017; Liu et al., 2017; Katzir et al., 2019; Lira et al., 2020; Park et al., 2020) propose to learn a deterministic generator, which maps each image of the source domain to a corresponding image of the target domain. These works often use a cycle consistency constraint, which enforces the generator to be bijective, thus preventing mode collapse. With this approach, the amount of possible target images one can generate per input image is often limited.

Other works (Huang et al., 2018; Lee et al., 2019) propose to view image-to-image translation as a style transfer problem, where the content is an image from the source domain, and the style is taken from the target domain. The style can be either a random noise from the desired style space or taken from some specific reference image in the target domain. By doing so, the number of possible target images that one can generate significantly increases. These works are multi-modal in the sense that a given image can be translated to multiple images in the target domain. This multi-modality can also be achieved in other approaches as shown by Nizan & Tal (2020).

While the aforementioned methods require training a generator for each pair of domains, some other works (Liu et al., 2019; Choi et al., 2020) combine style transfer with a training scheme that results in a single generator that can translate between any pair of domains or styles that appear during training. Moreover, Liu et al. (2019) shows that their method is capable of translating to styles that were unseen during training as long as the GAN was trained on closely-related styles.

In our work, we adopt the style transfer approach and use the training scheme that enables one generator to translate between multiple pairs of domains. While previous works focus on learning the translation between the desired domains, we also learn translations between *modalities* of the source domain, thus leveraging its richness. This makes our method multi-modal in the sense that it utilizes the modalities of the source domain for the training of the translation task. Although the apparent resemblance, the meaning of multi-modal (or cross-modal) in our work is fundamentally different than its meaning in MUNIT, in which multi-modality refers to the ability to translate a given image into multiple images in the target domain. Conversely, in our work, we refer to the latent modalities in the source domain.

Recently, it has been shown that the latent modalities of a dataset can assist in generating images, which belong to that dataset distribution (Liu et al., 2020; Sendik et al., 2020). The premise of these works is that real-world datasets cannot be well-represented using a uniform latent space, and information about their latent modalities helps to model the data distribution better. In our work, we exploit these modalities to improve the generator by training it to translate between them.

3 Method

BalaGAN aims at translating an image between the unpaired, rich source domain A, and a data-poor target domain B. To perform the translation, our method receives a source image and a reference image from the target domain. The source image is translated such that the output image appears to belong to the target domain. The training of our model consists of two steps: (i) finding k disjoint modalities in the source domain, where each modality is a set of images, denoted by A_i; (ii) training a single model to perform cross-translations among all pairs in $(A_1, ..., A_k, B)$, see Figure 1.

3.1 Finding Modalities

To find the modalities of a given domain, we train an encoder that yields a meaningful representation of the style for each image. Then, we cluster the representations of all source domain images, where each cluster represents a single modality.

We train our encoder following Chen et al. (2020), where contrastive loss is applied on a set of augmented images. Given a batch of images, we apply two randomly sampled sets of augmentations on each image. Then, we apply the encoder, and attract the result representations of augmented images if both were obtained from the same source image, and repel them otherwise. Choosing a set of augmentations that distort only content properties of the images, yields representations that are content agnostic and reflecting of the style. We use the normalized temperature-scaled cross-entropy loss (Chen et al., 2020; Sohn, 2016; Wu et al., 2018; Oord et al., 2018) to encourage a large cosine similarity between image representations with similar styles. As the dot product between such representations is small, spherical k-means allows for clustering images by their styles. We denote the clusters by $A_1, ..., A_k$, where k is chosen such that $|B| \geq |A|/k$ resulting in modalities which are relatively balanced. Analysis of different values of that k is given in Section 4.2.

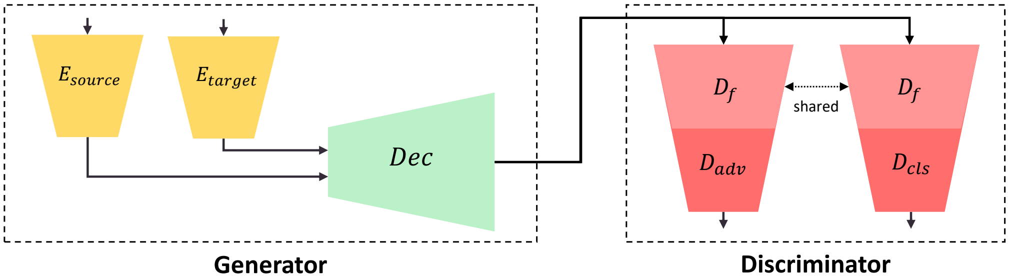


Figure 2: An illustration of BalaGAN's architecture.

3.2 Translation Network

Our translation network is a multi-class image-to-image translation network, where the classes $(A_1, ...A_k, B)$ are the clusters obtained above. The network cross-translates between all the $(k+1)^2$ pairs in $(A_1, ...A_k, B)$. The network's architecture and training procedure are built upon FUNIT (Liu et al., 2019). We train the translation network G, and a discriminator D in an adversarial manner. A high-level diagram of our architecture is shown in Figure 2.

G consists of source encoder E_s, target encoder E_t, and decoder F. Given a source image x, and a reference image y, the translated image is given by:

$$x' = G(x, y) = F(E_s(x), E_t(y)).$$

To train G, we sample two images from $(A_1, ..., A_k, B)$, a source image x, and a reference image y. The network receives these two images and generates a new image which resembles x on the pixel-wise level, but shares the same mode as y. At test time, we translate images from domain A to domain B by taking a source image from A and a reference image from B. Note that the trained network can translate any image from A without its cluster (modality) label.

Our discriminator consists of two sub-networks, which have shared weights in the initial layers, denoted by D_f. Each sub-network corresponds to a different task that the discriminator performs. The first sub-network, denoted by D_{adv}, aims to solve an adversarial task, in which it classifies each image to one of $(A_1, ..., A_k, B)$. That is, $D_{\text{adv}}(\cdot)$ is a $k+1$-dimensional vector with score for each modality. The translation network aims to confuse the discriminator, that is, given a source image x and a reference image y, G aims at making D_{adv} predict the modality of y for $G(x, y)$. For such a generated image, D_{adv} aims to predict any modality, but the modality of y, while for a real image it aims at predicting its correct modality. The second sub-network, denoted by D_{cls}, performs a classification task. This sub-network aims to predict the modality of each image, but here it is trained on the real images only. As shown in previous works (Chen et al., 2019), defining another task for the discriminator helps the stability of training, and eventually strengthens the generator. In Section 4.4 we show that this addition to the FUNIT architecture is significant for yielding better image translations.

Losses. We use a weighted combination of several objectives to train G and D. First, we utilize the Hinge version of the GAN loss for the adversarial loss (Liu et al., 2019; Lim & Ye, 2017; Miyato et al., 2018; Zhang et al., 2018; Brock et al., 2019). It is given by

$$\mathcal{L}_{\text{GAN}}(D) = E_x[\max(0, 1 - D_{\text{adv}}(x)_{m(x)})] + E_{x,y}[\max(0, 1 + D_{\text{adv}}(G(x, y))_{m(y)}],$$

$$\mathcal{L}_{\text{GAN}}(G) = -E_{x,y}[D_{\text{adv}}(G(x, y))_{m(y)}],$$

where $D_{\text{adv}}(\cdot)_i$ is the i-th index in the $k+1$-dimensional vector $D_{\text{adv}}(\cdot)$ and $m(x)$ is the modality of the image x. To encourage content-preservation of the source image and to help in preventing mode collapse we use a reconstruction loss. Additionally, to encourage the output image to resemble the reference image, we utilize the feature matching loss. They are given by

$$\mathcal{L}_{\text{R}}(G) = E_x[||x - G(x, x)||_1], \qquad \mathcal{L}_{\text{FM}}(G) = E_{x,y}[||D_f(G(x, y)) - D_f(y)||_1],$$

respectively. For the classification task of the discriminator, we use cross-entropy loss, defined by $\mathcal{L}_{\text{CE}}(D) = \text{CrossEntropy}(D_{\text{cls}}(x), \mathbf{1}_{m(x)})$, where $\mathbf{1}_{m(x)}$ is a one-hot vector that indicates the modality of the image x. Gradient penalty regularization term (Mescheder et al., 2018) is also utilized, given by $R_1(D) = E_x[||\nabla D_{\text{adv}}(x)||_2^2]$. The total optimization problem solved by our method is defined by

$$\min_D \mathcal{L}_{\text{GAN}}(D) + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}}(D) + \lambda_{\text{reg}}R_1(D), \qquad \min_G \mathcal{L}_{\text{GAN}}(G) + \lambda_{\text{R}}\mathcal{L}_{\text{R}}(G) + \lambda_{\text{F}}\mathcal{L}_{\text{FM}}(G).$$

Balanced setting. While the main motivation for the k-modal translation is for the imbalanced translation setting, our method also shows effectiveness in translation between two balanced domains, A and B. In such a setting, we split both A and B into modalities. Then, instead of defining the classes as $(A_1, ..., A_k, B)$, we define the classes to be $(A_1, ..., A_{k_s}, B_1, ..., B_{k_t})$ and train the translation network with all $(k_s + k_t)^2$ pairs.

4 Evaluation

We evaluate our cross-modal translation method in a series of experiments. We first show the effectiveness of our method in the imbalanced setting, by evaluating its performance when decreasing the number of images in the target domain. Next, we explore the influence of the number of modalities, k, on the result. Then, we show that our method can also be effective in the balanced setting. Finally, we perform an ablation study to compare our architecture with other alternative architectures and study the importance of finding effective modalities. To evaluate the results, we show a variety of visual examples, use the FID (Heusel et al., 2017) measurement, and perform a human perceptual study to validate the quality of the results obtained by our method compared to results of other leading methods.

Datasets. We use the CelebA dataset (Liu et al., 2015) and set the source and target domains to consist of 10,000 and 1000 images of women and men, respectively. We additionally use the Stanford Cars Dataset (Krause et al., 2013), and translate a range of different colored cars to red cars. There, the training set consists of 7500 non-red cars, and 500 red cars. From the AFHQ dataset (Choi et al., 2020) we take all the 4739 images of dogs as the source domain, and all the 5153 images of cats as the target domain. Furthermore, we use the Animal Face Dataset (AFD) (Liu et al., 2019) and set the source domain to be a mix of 16 breeds of dogs and the target domain to be a mix of three breeds of wolves. Our training set consists of 10,000 dog images and 1000 wolf images. It should be noted that among the above, the Animal Face Dataset is the most challenging due to the wide range of poses and image quality.

4.1 Effectiveness in the Imbalanced Setting

We compare our approach with other methods: CycleGAN (Zhu et al., 2017), U-GAT-IT (Kim et al., 2020), MUNIT (Huang et al., 2018), StarGAN2 (Choi et al., 2020), CUT (Park et al., 2020). We first train a number of methods on the AFD dataset. For our method, we used 40 modalities to train the translation network. Quantitative results are presented in Table 1.

BalaGAN	**CycleGAN**	**U-GAT-IT**	**MUNIT**	**StarGAN2**	**CUT**
60.88	77.8	97.16	83.38	211.77	108.64

Table 1: FID (↓) results of various image-to-image translation methods applied on AFD, translating dogs to wolves in the imbalanced setting. For BalaGAN we use 40 modalities.

For the above leading methods, we perform additional experiments over multiple datasets to show the effect of decreasing the number of training images in the target domain. Quantitative results over AFD and CelebA are presented in Table 2. As can be seen, CycleGAN and BalaGAN are the leading methods, and the image quality produced by BalaGAN is more stable as the size of the target domain decreases. Visual results are shown in Figure 3 for these two methods, and in Appendix C.1 for the other methods.

We further compare BalaGAN and CycleGAN through a human perceptual study, in which each user was asked to select the preferred image between images generated by these two methods. The images were generated by models that were trained using 1000 target domain images. 50 users participated in the survey, each answered ten random questions out of a pool of 200 questions for each dataset. As can be seen in Table 7b, BalaGAN outperforms CycleGAN on both datasets even though CycleGAN achieves lower FID for the women→men translation task.

Results for the Standford Cars dataset are presented in Figure 4. As can be seen, BalaGAN is almost agnostic to decrease in the size of the target domain, while CycleGAN is sensitive to such change.

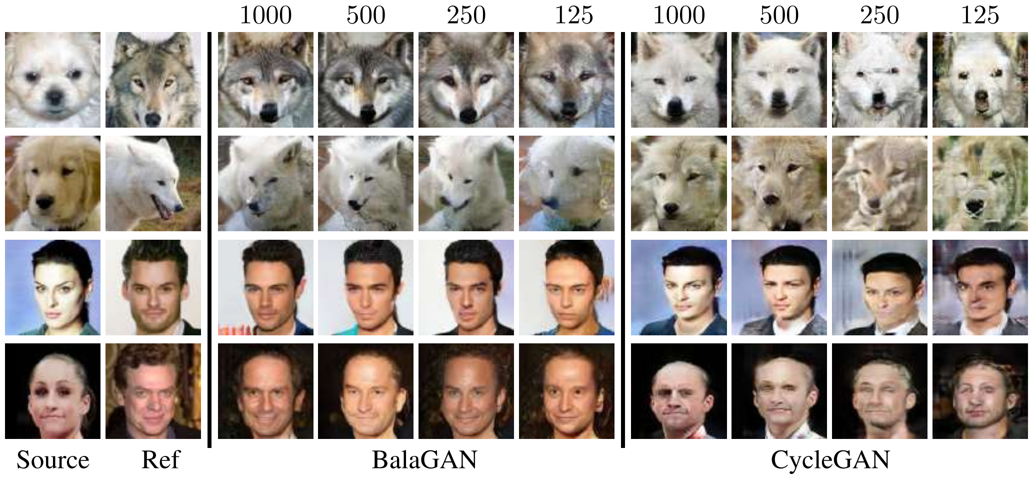


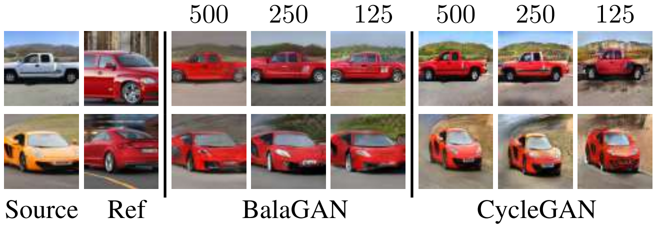
Figure 3: Applying CycleGAN and BalaGAN on the dog → wolf and woman → men translation tasks, by training with decreasing number of images in the target domain. The numbers above the table indicate the number of target domain images that were used for training.

|B|	**dogs → wolves**				**women → men**			
	BalaGAN	**CycleGAN**	**CUT**	**MUNIT**	**BalaGAN**	**CycleGAN**	**CUT**	**MUNIT**
1000	**60.88**	77.80	108.64	83.38	33.42	**28.33**	55.04	42.35
500	**72.46**	99.80	166.36	103.07	39.95	**38.59**	61.08	47.51
250	**102.35**	136.00	225.35	123.88	**38.99**	54.95	82.26	53.81
125	**157.67**	202.61	226.97	162.97	**49.42**	155.60	274.53	58.48

Table 2: FID results (↓) applied on AFD and CelebA datasets in the imbalanced setting. $|B|$ denotes the number of images in the target domain that were used during training.

4.2 The Influence Of k

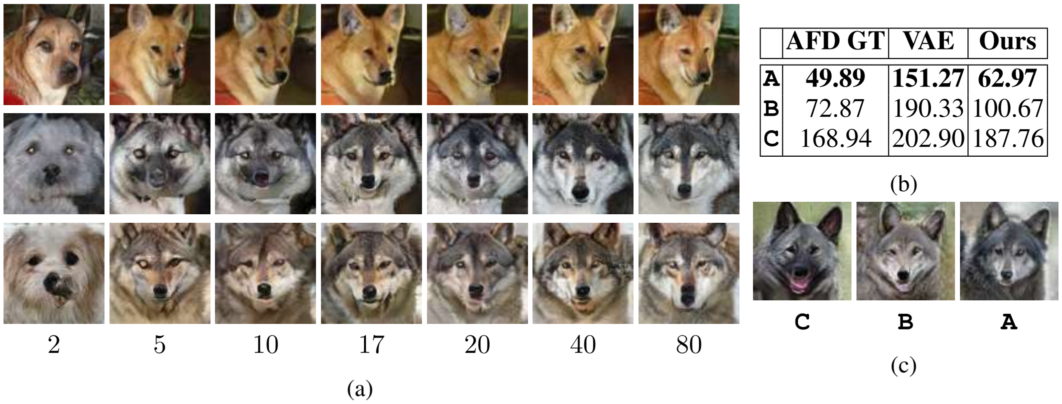
The number of modalities that our translation network is trained on, $k + 1$, is an important factor for the success of our method. For $k = 1$, our method is reduced to the common setting of image-to-image translation, and as we increase k, our network is enforced to train and learn more translation tasks, resulting in more accurate translation. Here we show that the value of k influences the quality



|B|	**car → red-car**	
	BalaGAN	**CycleGAN**
500	37.02	**33.38**
250	40.38	40.46
125	**40.25**	51.17

Figure 4: Visual (left) and FID(↓) (right) results of CycleGAN and BalaGAN applied on the car → red-car translation task, by training with decreasing number of images in the target domain.

of the generated images. Visual results that were obtained on the dog → wolf translation task are shown in Figure 5a and quantitative results are provided in Figure 7d. As can be seen, as k increases, FID decreases, i.e., the images quality is improved. Note, however, that once k goes beyond 16, the number of dog breeds, the improvement of the results is rather moderate.



	AFD GT	VAE	Ours
A	**49.89**	**151.27**	**62.97**
B	72.87	190.33	100.67
C	168.94	202.90	187.76

(b)

Figure 5: (a) Results of BalaGAN applied with varying values of k. Below each column we specify the number of modalities that the translation network was trained on, that is $k+1$. (b) FID(↓) of our ablation study applied on the dog → wolf translation task with $k = 16$ which is the number of dogs' breeds. Rows and columns notation are explained in 4.4 (c) Visual results of ablation study.

4.3 Effectiveness in the Balanced Setting

Here we present results on a balanced dataset. We choose the AFHQ dataset, translating dogs to cats. We train BalaGAN using latent modalities extracted in both the source and target domain. For this dataset, we extracted 30 modalities in each domain. We compare our method with five strong baseline methods: CycleGAN (Zhu et al., 2017), CUT (Park et al., 2020), GANHopper (Lira et al., 2020), StarGAN2 (Choi et al., 2020), and MUNIT (Huang et al., 2018). We denote the StarGAN2 that is trained on the two domains as StarGAN2^1, and StarGAN2 that is trained to translate between each pair of the 60 modalities that we find as StarGAN2^{30}. For MUNIT, we show results when the style is taken from a reference image (denoted by MUNIT^r), and from a random noise vector (denoted by MUNIT^n). Figure 6 shows a random sample of results from this comparison, and in Table 3 we present a quantitative comparison. As can be observed, our method outperforms other methods both visually and quantitatively.

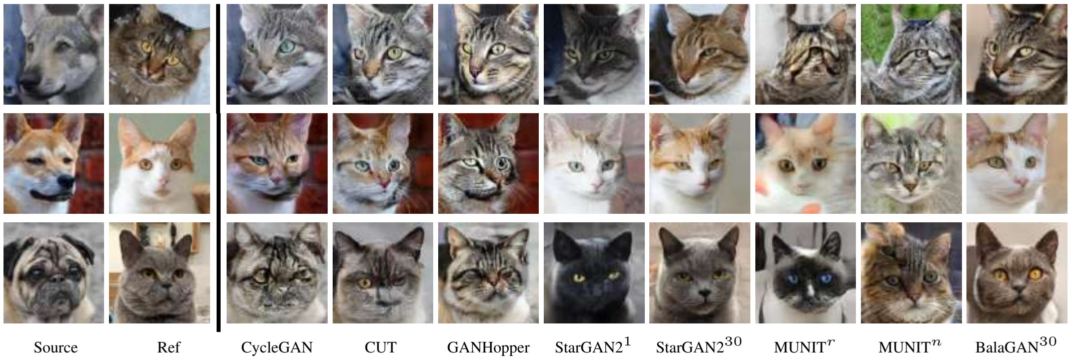
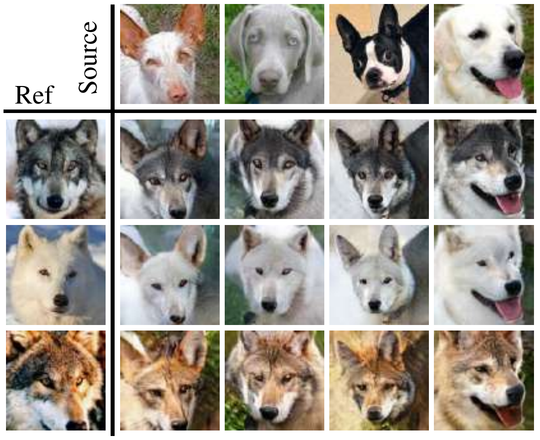


Figure 6: Various methods applied on AFHQ dataset, which is balanced, to translate dogs to cats. The super-index denotes k. Additional results are shown in Appendix C.2.

As the leading methods according to the FID measure are BalaGAN and StarGAN2, we further compared them through a human perceptual study. Similarly to the imbalanced user study, each user answered 10 random questions out of a pool of 200 questions. Here, the user was asked to

CycleGAN	CUT	GANHopper	StarGAN2^1	StarGAN2^{30}	MUNIT^r	MUNIT^n	BalaGAN
29.98	27.37	33.79	29.56	25.89	35.80	27.11	**19.21**

Table 3: FID (↓) results of applying various image-to-image translation methods over AFHQ dataset.



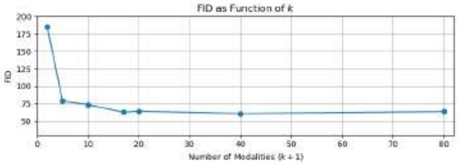
(a)

Task	BalaGAN	CycleGAN
dogs → wolfs	**83.3**	16.7
women → men	**66.4**	33.6

(b)

BalaGAN	StarGAN^{30}	StarGAN^1
43.9	27.4	28.7

(c)



(d)

Figure 7: (a) BalaGAN on the AFD trained on 1000 wolves using 40 modalities. (b) Human perceptual study in the imbalanced setting. We present the percentage of users that chose the corresponding image as the preferred one. (c) Users preferences for the AFHQ dataset in a balanced setting. (d) FID (↓) of our method applied on AFD in an imbalanced setting. The number of modalities is $k + 1$.

choose between images of BalaGAN, StarGAN^{30} and StarGAN^1. As observed in Table 7c, most users chose images of BalaGAN, where the scores of StarGAN^{30}, and StarGAN^1 are similar.

4.4 Diversity and Ablation Study

The diversity of generated images that can be achieved by our method, is shown in Figure 7a (see additional results in Appendix C.3). We additionally perform an ablation study, in which we change the translation network and the decomposition of the source domain. For the ablation of the translation network, let **A** denote our BalaGAN method, then (i) in **B** we removed the D_{cls} loss, and (ii) in **C**, we additionally do not use the target domain images during training. Note, that the setting in **C** degenerates into FUNIT (Liu et al., 2019). For the ablation of the source's decomposition, let AFD GT denote the dogs' breeds ground truth class labels and VAE denotes a variational autoencoder that replaces our encoder. The results presented in Table 5b and Figure 5c show that D_{cls} significantly improves the architecture of FUNIT, even in a transductive setting.

We explore the robustness of our method by comparing the results of two variations of StarGAN2, one trained on two domains and one trained on the learned modalities, denoted by StarGAN2^1 and StarGAN2^{30} respectively. The results are shown in Figure 6 and Table 3. As one can see, training StarGAN2 to translate between modalities improves the network's ability to translate between the two domains. Therefore, we conclude that the benefit of training on modalities is not specific to our architecture, and can be utilized by other multi-class image-to-image translation methods.

5 Conclusion

We have presented an image-to-image translation technique that leverages latent modes in the source and target domains. The technique was designed to alleviate the problems associated with the imbalanced setting, where the target domain is poor. The key idea is to convert the imbalanced setting to a balanced one, where the network is trained to translate between all pairs of modes, including the target one. We have shown that the balanced setting leads to better translation than strong baselines. We further showed that analyzing and translating at the mode-level, can benefit also in a balanced setting, where both the source and target domains are split and the translator is trained on all pairs. In the future, we would like to use our technique to re-balance training sets and show that downstream applications, like object classification and detection, can benefit from the re-balancing operation. We believe our work to be a step in the direction of analyzing domain distributions and learning their latent modes, and would like to reason and apply this idea on a wider range of problems beyond image-to-image translation.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in neural information processing systems, pp. 6626–6637, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In ECCV, 2018.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734, 2017.

Oren Katzir, Dani Lischinski, and Daniel Cohen-Or. Cross-domain cascaded deep feature translation, 2019.

Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=BJlZ5ySKPH.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.

Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation viadisentangled representations. arXiv preprint arXiv:1905.01270, 2019.

J. H. Lim and J. C. Ye. Geometric gan. ArXiv, abs/1705.02894, 2017.

Wallace Lira, Johannes Merz, Daniel Ritchie, Daniel Cohen-Or, and Hao Zhang. Ganhopper: Multi-hop gan for unsupervised image-to-image translation, 2020.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems 30, pp. 700–708. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In IEEE International Conference on Computer Vision (ICCV), 2019.

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14286–14295, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

Lars M. Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In ICML, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7860–7869, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In European Conference on Computer Vision, 2020.

Omry Sendik, Dani Lischinski, and Daniel Cohen-Or. Unsupervised k-modal styled content generation. ACM Trans. Graph., 39(4), July 2020. ISSN 0730-0301. doi: 10.1145/3386569.3392454. URL https://doi.org/10.1145/3386569.3392454.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), Advances in Neural Information Processing Systems 29, pp. 1857–1865. Curran Associates, Inc., 2016.

A.K Subramanian. Pytorch-vae. https://github.com/AntixK/PyTorch-VAE, 2020.

Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. arXiv:1805.08318, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Computer Vision (ICCV), 2017 IEEE International Conference on, 2017.

A Additional Details

A.1 Implementation Details

Finding Modalities. For finding the modalities, we use ResNet-18 as the encoder and train it from scratch. We train the encoder over all training samples in the source and target domains. Input images are resized to 128×128, and the embedding dimension is set to 256. We use Adam optimizer with a learning rate of $3e^{-4}$ and weight decay of $1e^{-6}$. To cluster the images into modalities, we use the spherical k-means implementation of Johnson et al. (2017). We cluster only the source domain images, while keeping the target domain as its own cluster. For the balanced case, we cluster the source and target domains separately.

Translation Network. For the translation network, we built our model based on FUNIT implementation. The architectures of D_{adv} and D_{cls} are the same, and these networks share all the layers besides the last two convolution blocks. We train the translation network to generate 128×128 images, using a batch size of 10 and 150,000 iterations for the CelebA dataset, and 100,000 iterations for the other datasets.

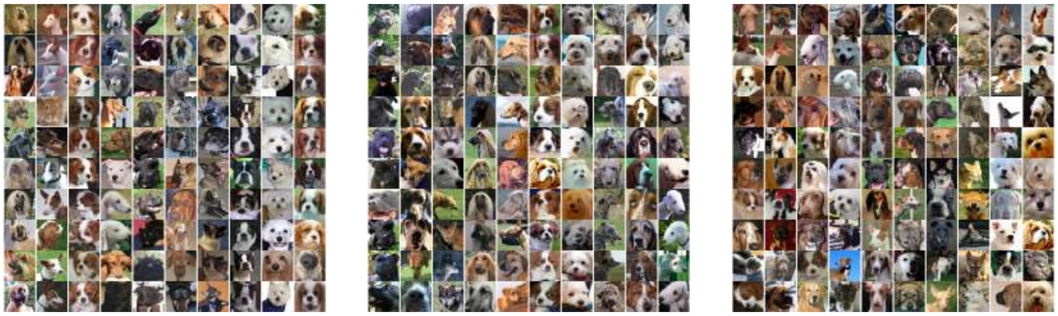
A.2 Ablation Study

As Explained in 4.4, we explore the influence of the modalities on the performance of the translation network. To do that, we (i) define the modalities according to the ground-truth dog breeds obtained from ImageNet labels, and (ii) train a variational autoencoder, and cluster the images embeddings obtained by the encoder. We define the modalities to be the result clusters. We use the variational autoencoder implemented by Subramanian (2020).

B Result Modalities

In this section we present the modalities that were extracted by our method and discuss the augmentations used for each dataset.

AFD. For this dataset, the augmentations that we use are: crop, horizontal-flip, color-distortion, gray-scale, and blur. We show the modalities that are obtained by our method compared to the modalities that are obtained by the variational-autoencoder in Figure 8. Here, 10 modalities are randomly sampled from the results of each method.



BalaGAN 40 Modalities

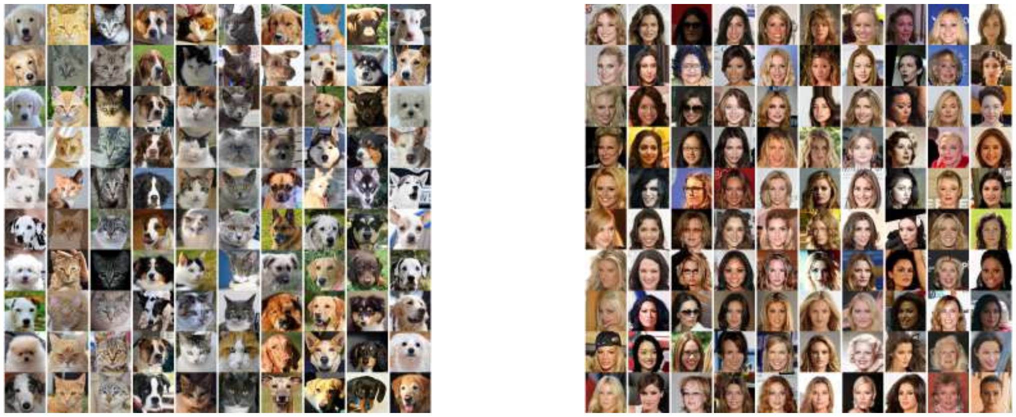
BalaGAN 17 Modalities

VAE Modalities

Figure 8: Modalities that are extracted by clustering the representations obtained by BalaGAN's encoder, compared to those that obtained by a VAE.

AFHQ. Here we use the same augmentations as in AFD, and split both the source and target domains into 30 modalities each. The obtained modalities are presented in Figure 9a.

CelebA. Here we used the same augmentations as in AFD, and split the source domain into 30 modalities. The obtained modalities are presented in Figure 9b.

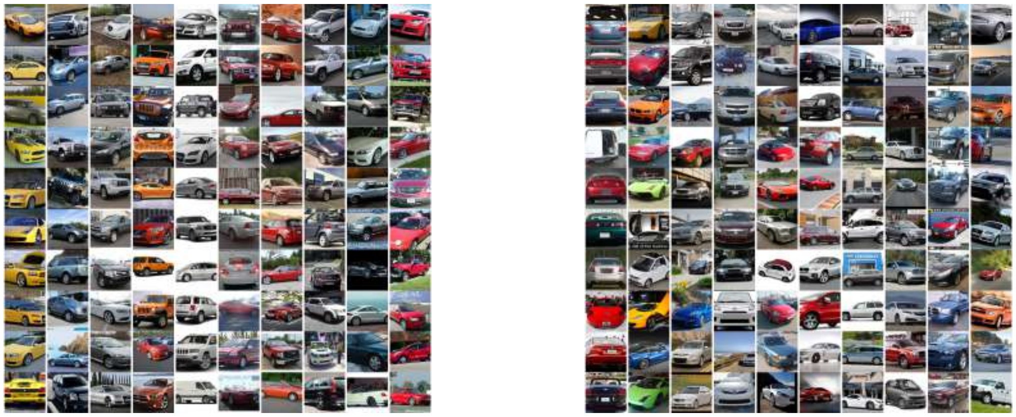


(a)

(b)

Figure 9: Modalities that were obtained by BalaGAN for (a) AFHQ dataset and (b) CelebA dataset.

Stanford Cars Dataset. To demonstrate the effect of the augmentations chosen to train the encoder, we show clusters that were obtained by training the encoder with two different augmentations sets. One set of augmentations yields clusters that are associated with *style* while the other set yields *content-related* clusters. For finding style clusters we use $\{\text{crop}, \text{horizontal flip}, \text{shuffle}, \text{gray-scale}, \text{blur}\}$ and for the content clusters we use $\{\text{crop}, \text{color-distortion}, \text{gray-scale}, \text{blur}\}$. The results are shown in Figure 10.



(a) Style Clusters

(b) Content Clusters

Figure 10: Clusters that were obtained by training the encoder with two different sets of augmentations. Each column corresponds to a cluster.

C Additional Results

In this section we show additional visual results. We first show a comparison of our method with other methods when decreasing the target domain size (Figures 11 and 12). Then, we show a comparison of our method with other methods in the balanced setting (Figure 13). Finally, we show that our method can generate diverse images in the target domain, which enable to enrich the target domain. Diverse images can be obtained by changing the source and reference images (Figures 15, 16, 17), and further by *interpolating* between images in the latent space (Figure 14).

C.1 Decreasing Target Domain Size

Here we show additional results of various methods that are trained over a training data with decreasing number of images in the target domain. For each method, we used the official implementation, and used the default training configuration with images resized to 128×128.

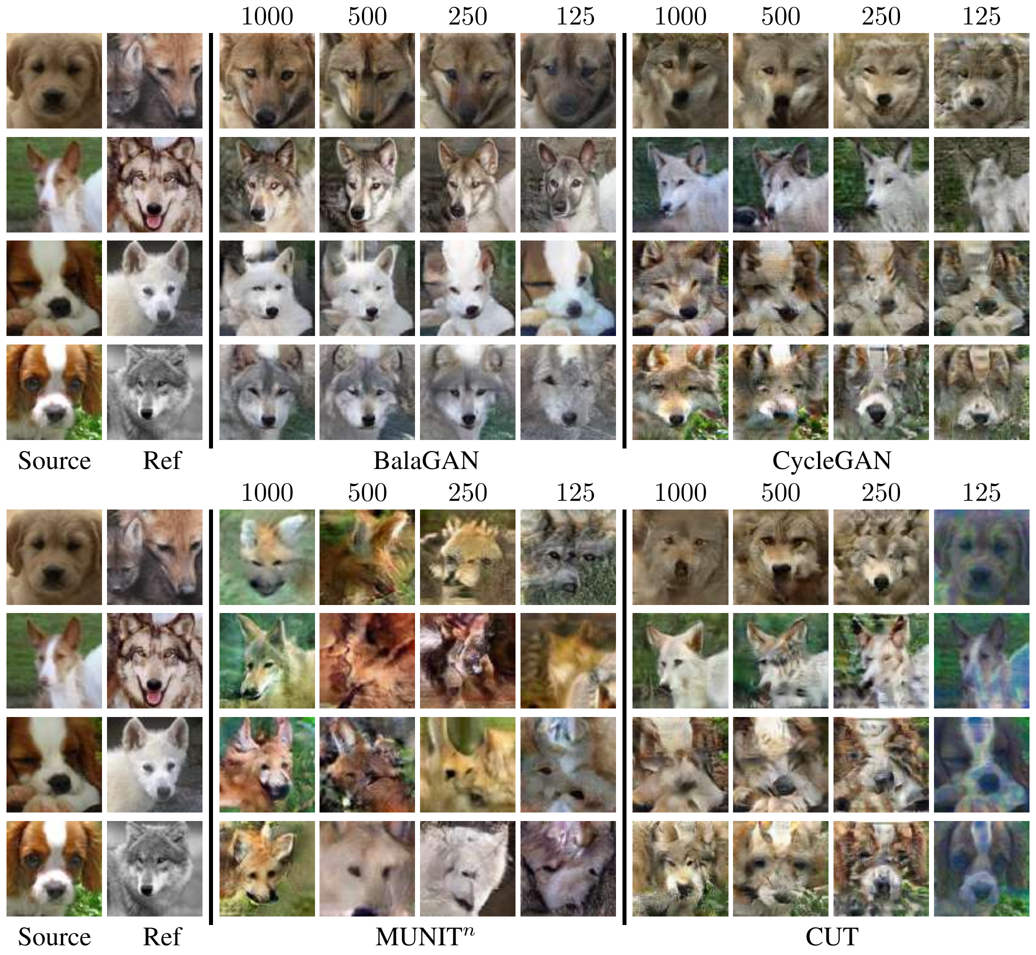


Figure 11: Applying CycleGAN, MUNIT^n, CUT and BalaGAN on the dog → wolf translation task, by training with decreasing number of images in the target domain. The numbers above the table indicate the number of target domain images that were used for training. As can be seen, BalaGAN achieves the best results for the imbalanced setting. The results of CUT resemble those of CycleGAN, but the performance decline in CUT is more significant. MUNIT struggles in learning the varied wolves distribution out of a small target domain.

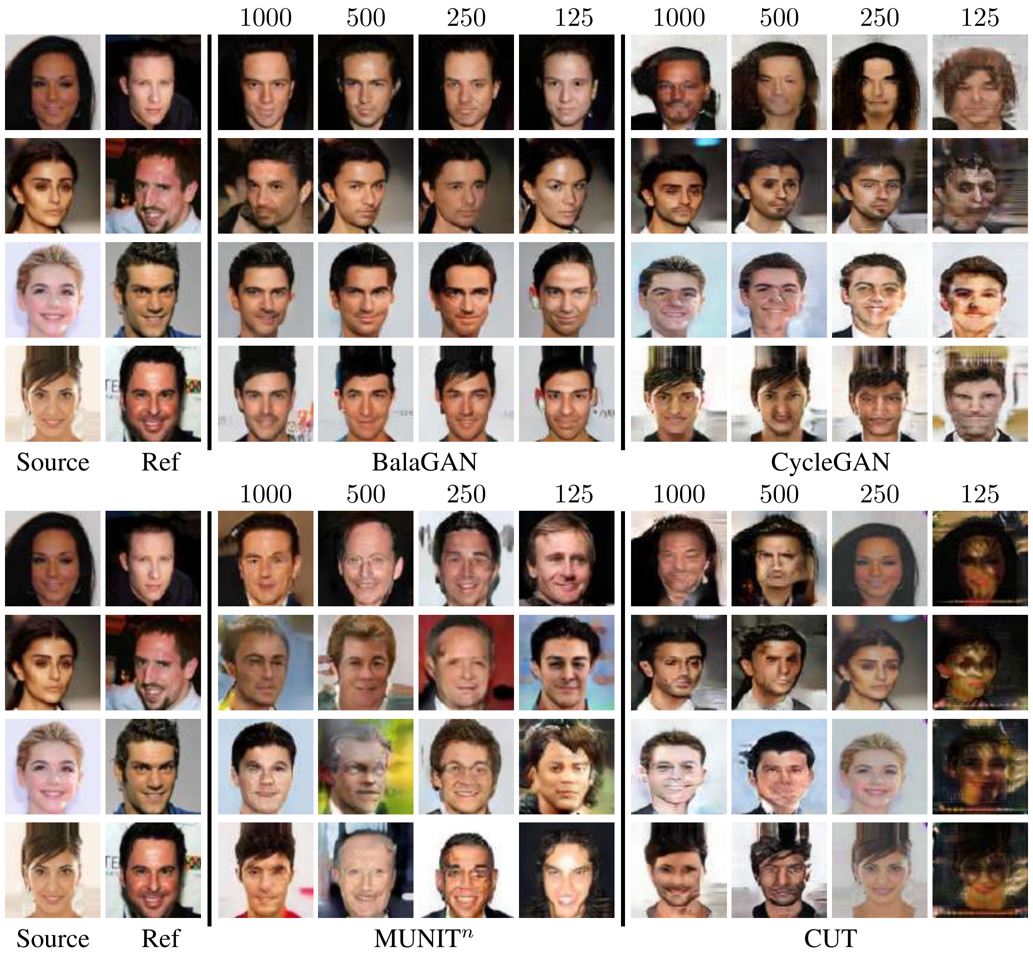


Figure 12: Applying CycleGAN, MUNIT^n, CUT and BalaGAN on the women → men translation task, by training with decreasing number of images in the target domain. The numbers above the table indicate the number of target domain images that were used for training. As can be seen, BalaGAN is almost agnostic to decrease in the size of the target domain. While MUNIT struggled in learning the varied wolves distribution, here it produces better results since the distribution of men's faces is not as varied.

C.2 Balanced Setting

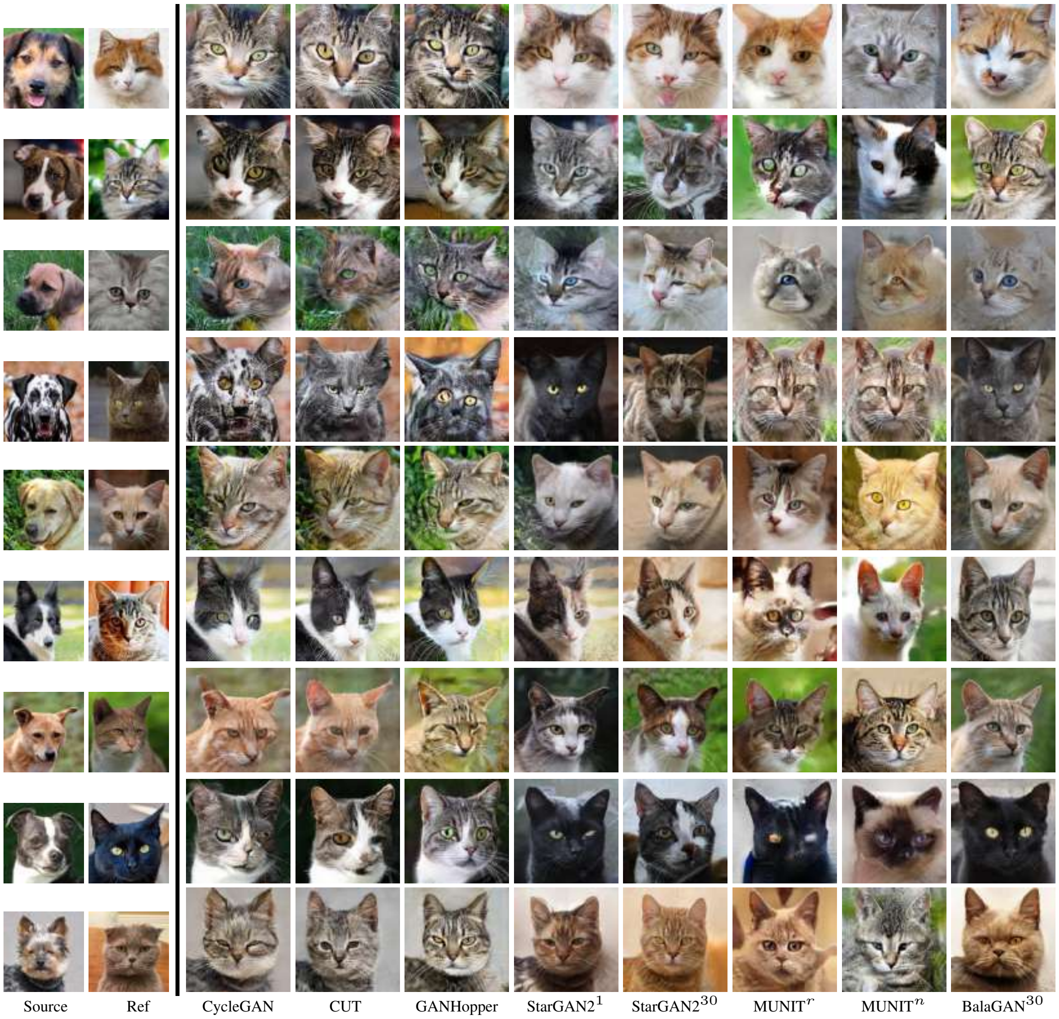


Figure 13: Results of various methods applied on AFHQ in the balanced setting.

C.3 BalaGAN Additional Results

Here we show additional results of BalaGAN. For AFD and CelebA we trained our method using 1000 images in the target domain. For AFHQ we used the balanced setting of our approach.

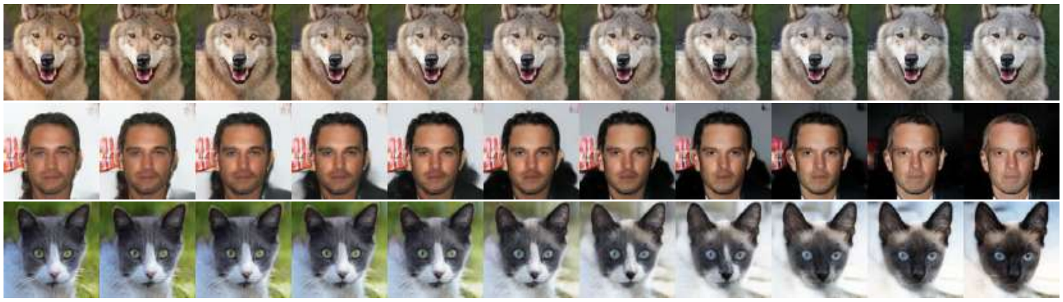


Figure 14: As our translation network is applied in latent space, it is possible to interpolate between two given reference images. This significantly increases the ability of enriching the target domain.

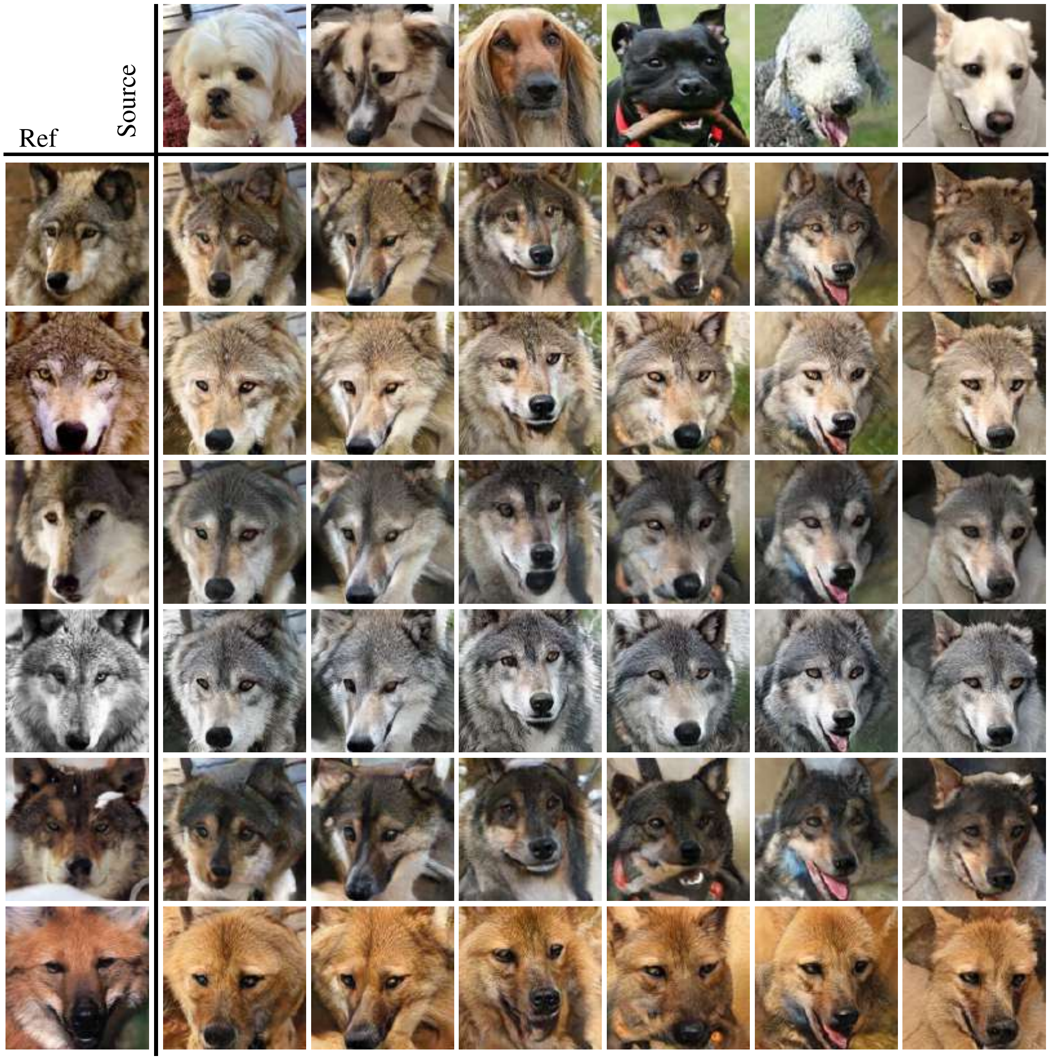


Figure 15: BalaGAN applied on the dogs→wolves translation task. We trained our method over 10,000 dogs and 1000 wolves, using 40 modalities.

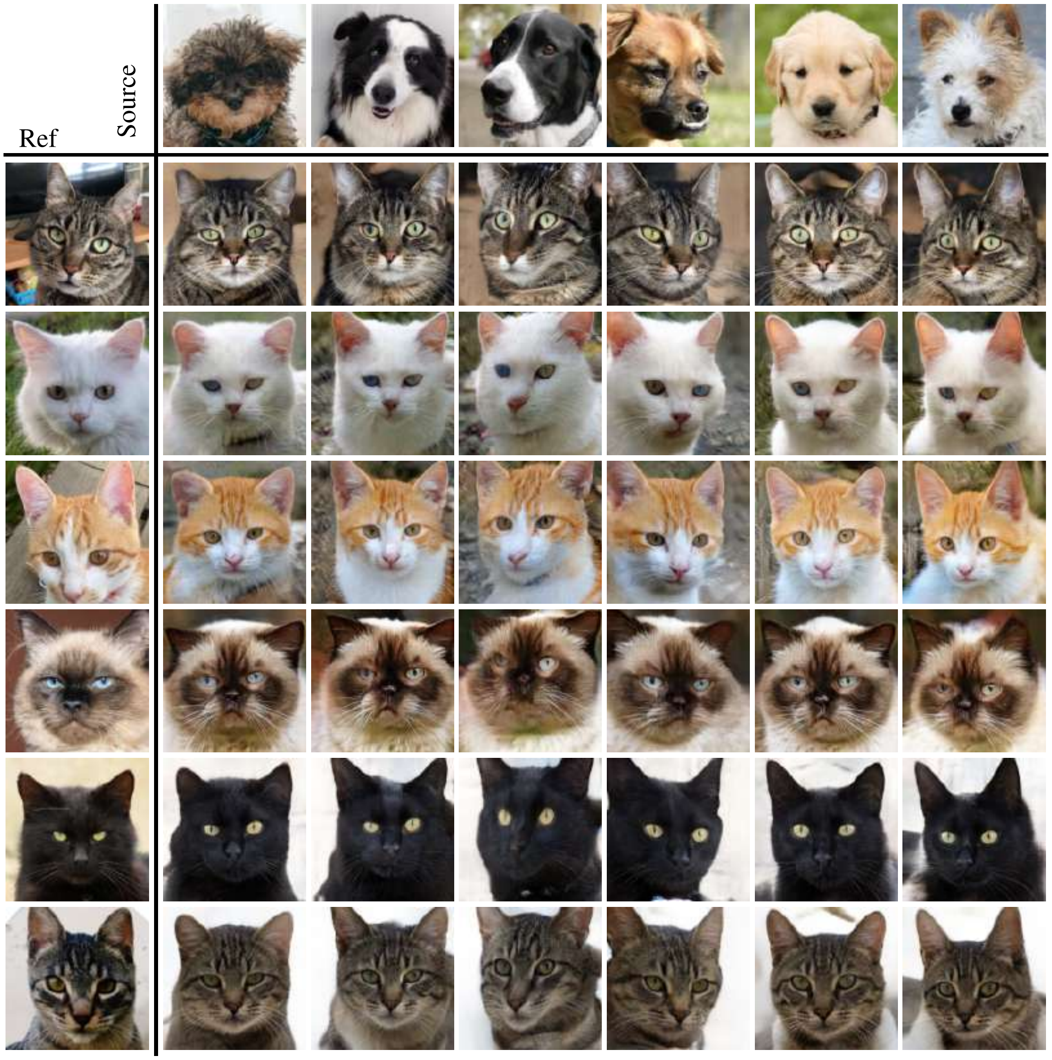


Figure 16: BalaGAN applied on the balanced dogs→cats translation task. We decomposed both the source and target domains into 30 modalities.

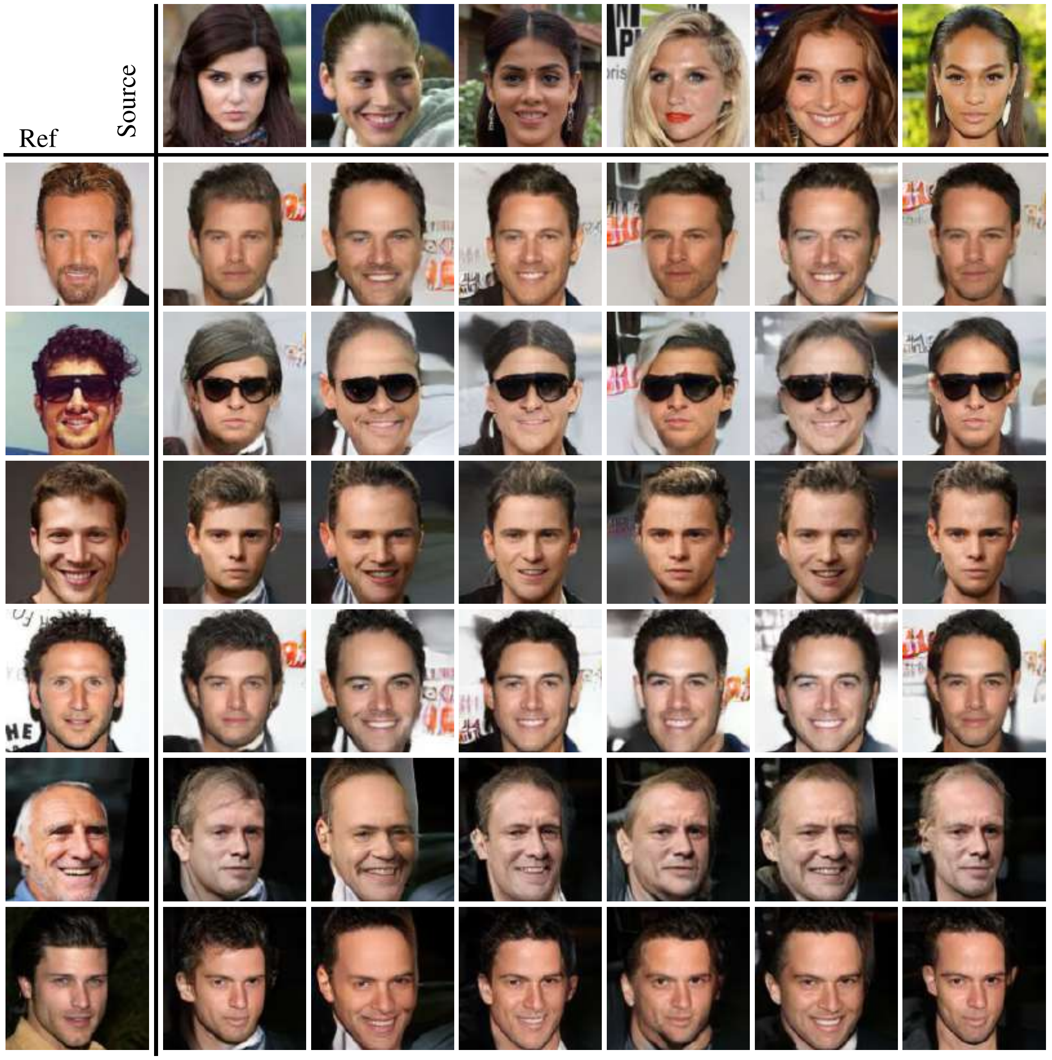


Figure 17: BalaGAN applied on the women→men translation task. We trained our method over 10,000 women and 1000 men, using 30 modalities.