CrossNet: Latent Cross-Consistency for Unpaired Image Translation

Omry Sendik
Tel Aviv University
omrysendik@gmail.com

Dani Lischinski
The Hebrew University of Jerusalem

Daniel Cohen-Or
Tel Aviv University

Figure 1: Given two unpaired sets of images, we train a model to perform translation between the two sets. Here we show, from left to right, our results on changing a specular material to diffuse, enhancing a mobile phone image to look like one taken by a Digital SLR camera and foreground extraction.

Abstract

Recent GAN-based architectures have been able to deliver impressive performance on the general task of image-to-image translation. In particular, it was shown that a wide variety of image translation operators may be learned from two image sets, containing images from two different domains, without establishing an explicit pairing between the images. This was made possible by introducing clever regularizers to overcome the under-constrained nature of the unpaired translation problem.

In this work, we introduce a novel architecture for unpaired image translation, and explore several new regularizers enabled by it. Specifically, our architecture comprises a pair of GANs, as well as a pair of translators between their respective latent spaces. These cross-translators *enable us to impose several regularizing constraints on the learnt image translation operator, collectively referred to as* latent cross-consistency. *Our results show that our proposed architecture and latent cross-consistency constraints are able to outperform the existing state-of-the-art on a variety of image translation tasks.*

1. Introduction

Many useful operations on images may be cast as an *image translation* task. These include style transfer, image colorization, automatic tone mapping and many more. Several such operations are demonstrated in Figure 1. While each of these operations may be carried out by a carefully-designed task-specific operator, in many cases, the abundance of digital images along with the demonstrated effectiveness of deep learning architectures, makes a data-driven approach feasible and attractive.

A straightforward supervised approach is to train a deep network to perform the task using a large number of pairs of images, before and after the translation [9]. However, collecting a large training set consisting of paired images is often prohibitively expensive or infeasible.

Alternatively, it has been demonstrated that an image translation operator, may also be learned from two image sets, containing images from two domains A and B, respectively, without establishing an explicit pairing between images in the two sets [28, 26, 16]. This is accomplished using generative adversarial networks (GANs) [4].

This latter approach is more attractive, as it requires much weaker supervision, however, this comes at the cost of making the translation problem highly under-constrained. In particular, a meaningful pairing is not guaranteed, as there are many pairings that are able to yield the desired distribution of translated images. Furthermore, undesirable phenomena, such as mode collapse, may arise when attempting to train the translation GAN [4].

To address these issues, existing GAN-based approaches for unpaired image translation [28, 26], train two GANs. One GAN maps images from domain A to domain B, and a second one operates in the opposite direction (from B to A). Furthermore, a strong regularization is imposed in the form

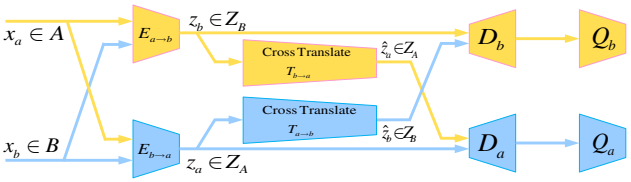

Figure 2: Our architecture uses two GANs that learn an image translation operator from two unpaired sets of images, A and B. By introducing a pair of *cross-translators* between the latent spaces (Z_B and Z_A) of the two Encoder-Decoder generators, we enable several novel latent cross-consistency constraints. Note that the only components used to perform the translation at test time are the encoders E and the decoders D, while the other components (cross-translators T and discriminators Q), as well as the crossing data paths, are only used by the training phase.

of the *cycle consistency loss*, which ensures that concatenating the two translators roughly reconstructs the original image. Note that the cycle consistency loss is measured using a pixelwise metric (L_1) in the original image domains.

In this work, we introduce a novel architecture for unpaired image translation, and explore several new regularizers enabled by it. Our architecture also comprises a pair of GANs (for $A \to B$ and $B \to A$ translation), but we also add a pair of translators between the latent spaces of their respective generators, Z_B and Z_A, as shown in Figure 2. These *cross-translators* enable us to impose several regularizing constraints on the learnt image translation operator, collectively referred to as *latent cross-consistency*. Intuitively, regularizing the latent spaces is a powerful yet flexible approach: the latent representation computed by the GAN's generator captures the translation task's most pertinent information, while the original input representation (pixels) contains much additional irrelevant information.

We demonstrate the competence of the proposed architecture and latent cross-consistency, as well as several additional loss terms, through an ablation study and comparisons with existing approaches. We show our competitive advantages vs. the existing state-of-the-art on tasks such as translating between specular and diffuse objects, translating mobile phone photos to DSLR-like quality and foreground extraction. Figure 1 demonstrates some of our results.

2. Related work

2.1. Unpaired image-to-image translation

2017 was a year with multiple breakthroughs in unpaired image-to-image translation. Taigman et al. [23] kicked off by proposing an unsupervised formulation employing GANs for transfer between two unpaired domains, demonstrating transfer of SVHN images to MNIST ones and of face photos from the Facescrub dataset to emojis.

Two seminal works which achieved great success in unpaired image-to-image translation are CycleGAN [28] and DualGAN [26]. Both proposed to regularize the training procedure by requiring a bijection, enforcing the translation from source to target domain and back to the source to return to the same starting point. Such a constraint yields a meaningful mapping between the two domains. Furthermore, since bijection cannot be achieved in the case of mode collapse, it thus prevents it.

Dong et al. [3] trained a conditional GAN to learn shared global features from two image domains, then followed by synthesis of plausible images in either domain from a noise vector conditioned on a class/domain label. To enable image-to-image translation, they separately train an encoder to learn a mapping from an image to its latent code, which would serve as the noise input to the conditional GAN to generate a target image.

Choi et al. [2] proposed StarGAN, a network that learns the mappings among multiple domains using only a single generator and a discriminator. Kim et al. [11] tackled the lack of image pairing in the image-to-image translation setting through a model based on two different GANs coupled together. Each GAN ensures that its generative function can map its domain to that of its counterpart. Since their method discovers relations between different domains, it may be leveraged to successfully transfer style.

A very different recent approach is NAM [6], which relies on having a high quality pre-trained unsupervised generative model for the source domain. If such a generator is available, a generative model needs to be trained only once per target dataset, and can thus be used to map to many target domains without adversarial generative training.

In this work we also address unpaired image-to-image translation. Contemporary approaches and some of those mentioned above tackle this problem by imposing constraints formulated in the image domain. Our approach consists of novel regularizers operating *across* the two latent spaces. Through the introduction of a unique architecture which enables a strong coupling between a pair of latent spaces, we are able to define a set of losses which are domain-agnostic. Additionally, a benefit of our architecture is that it enables multiple regularizers, which together push the trained outcome to a more stable final result. We stress that this is different from contemporary approaches relying on image domain losses, which make use of one or two losses (usually an identity or cycle-consistency loss).

2.2. Latent space regularization

Motivated by the fact that image-to-image translation aims at learning a joint distribution of images from the source and target domains, by using images from the marginal distributions in individual domains, Liu et al. [15]

made a shared-latent space assumption, and devised an architecture which maps images from both domains to the same latent space. By sharing weight parameters corresponding to high level semantics in both the encoder and decoder networks, the coupled GANs are enforced to interpret these image semantics in the same way. Additionally, VeeGAN [22] also addressed mode collapse by imposing latent space constraints. In their work, a reconstructor network reverses the action of the generator through an architecture which specifies a loss function over the latent space.

The two works mentioned above attempt to translate an image from a source domain A to a single target domain B. The scheme by which they achieve this limits their ability to extend to translating an input image to multiple domains at once. Armed with this realization, Huang et al. [7] proposed a multimodal unpaired image-to-image translation (MUNIT) framework. Achieving this involved decoupling the latent space into content and style, under the assumption that what differs between target domains is the style alone.

Tackling the lack of diversity of the results, Lee et al. [13] proposed Diversifying Image-to-Image Translation (DRIT) by embedding images onto both a domain-invariant content space and a domain-specific attribute space.

Mejjati et al. [18] proposed AGGAN, improving image to image translation results by adding attention guidance, showing that their approach focuses on relevant regions in the image without requiring supervision.

In the entirety of these architectures there is no path within the network graph, which enables formulating losses that constrain both latent spaces at once. For this reason, we dissect the common GAN architecture, and propose a path between encoders and decoders from cross (opposite) domains. Our architecture thus consists of a pair of GANs, but in addition, we couple each generator with a translator between latent spaces. The addition of the translators opens up not only the ability to enforce bijection constraints in latent space but more intriguing losses, which further constrain the problem, leading to better translations.

3. Cross Consistency Constraints

Architectures such as CycleGAN [28] or DualGAN [26] are able to accomplish unpaired image-to-image translation by imposing consistency constraints in the original image domains A and B. Thus, their constraints operate on the original, pixel-based, image representations, which contain much information that is irrelevant to the translation task at hand. However, it is well-known that in a properly trained Encoder-Decoder architecture, the latent space contains a distillation of the features that are the most relevant and pertinent to the task. To demonstrate this, consider the Horse-to-Zebra translation task, for example. The top row of Figure 3 visualizes the latent code of CycleGAN's Horse-to-Zebra Encoder-Decoder generator, where we can see that the latent code already contains zebra-specific information, such as the locations and shapes of the zebra stripes. In a manner of speaking, the generator's encoder has already planted "all the makings of a zebra" into the latent code, leaving the decoder with the relatively simpler task of transforming it back into the image domain. Similarly, the second row of Figure 3 demonstrates a case where the zebra-specific features are embedded in the wrong spatial regions, yielding a failed translation result.

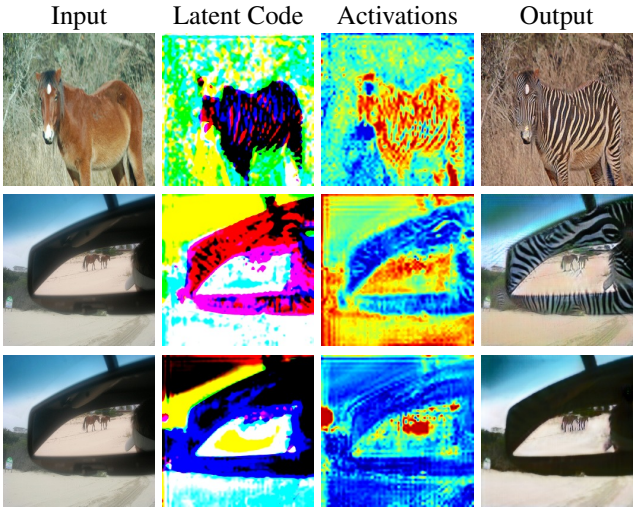

Figure 3: Latent space visualization for two Horse-to-Zebra translation examples. The second column visualizes the latent space of the generator by using PCA to reduce the 256 channels of the latent space to three, mapped to RGB. Alternatively, the third column shows the magnitude of the 256-dimensional feature vector at each latent space neuron. Note that the latent space in these examples indicates the positions and shapes of the zebra stripes in the resulting translated image. The two upper rows show the latent space and results of CycleGAN. The bottom row shows the latent space and result in our approach, where the encoder does not attempt to texture improper regions, thanks to its training with our latent cross-consistency losses.

The above observation motivates us to explore an architecture that enables imposing consistency constraints on and between the latent spaces. In our architecture, the latent spaces of the two Encoder-Decoder generators (from A to B and vice versa) are coupled via a pair of *cross-translators*. This creates additional paths through which data can flow during training, enabling several novel latent cross-consistency constraints in the training stage, as described below. The bottom row of Figure 3 shows an example where our encoder, trained with these constraints, avoids the incorrect embedding of the zebra specific features.

3.1. Architecture

Armed with the motivation to impose regularizations in latent space, we propose an architecture which links between the latent spaces of the generators (from domain A to B and vice versa), thus enabling a variety of latent consistency constraints. Our architecture is shown in Figure 2.

The architecture consists of a pair of Encoder-Decoder generators, which we denote by $(E_{a\to b}, D_b)$, and $(E_{b\to a}, D_a)$. The encoder $E_{a\to b}$ encodes an image $x_a \in A$ to a latent code denoted by $z_b \in Z_B$, while D_b decodes it to an output image $\hat{x}_b$. Similarly, the encoder $E_{b\to a}$ encodes an image $x_b \in B$ to a latent code denoted by $z_a \in Z_A$, while D_a decodes it to an output image $\hat{x}_a$. The discriminators, Q_b and Q_a, attempt to determine whether or not an input image from domains B or A, respectively, is real or fake. The novel part of our architecture is the addition of two *cross-translators*, $T_{b\to a}$ and $T_{a\to b}$, shown in Figure 2. Each translator is trained to transform the latent codes of one generator into those of the other, namely, from z_b to $\hat{z}_a$ and from z_a to $\hat{z}_b$, respectively. By adding the two cross-translators to our architecture, several additional paths, through which data may flow during training, become possible, paving the way for new consistency constraints. In this work, we present three novel latent cross-consistency losses, which are shown to conjoin to produce superior results on a variety of image translation tasks.

Note that the cross-translators T and the discriminators Q are only used at train time. At test time, the only components used for translating images are the encoders E and the decoders D.

3.2. Latent Cross-Identity Loss

In order to train the cross-translators $T_{a\to b}$ and $T_{b\to a}$, we require that an image x_a fed into the encoder $E_{a\to b}$ should be reconstructable by the decoder of the dual generator, D_a, after translation of its latent code by $T_{b\to a}$. A symmetric requirement is imposed on the translator $T_{a\to b}$. These two requirements are formulated as the *latent cross-identity loss*:

$$\begin{aligned}\mathcal{L}_{ZId} = \mathbb{E}_{x_a\sim p_{data}(A)}[\|D_a(T_{b\to a}(E_{a\to b}(x_a))) - x_a\|_1]+\\ \mathbb{E}_{x_b\sim p_{data}(B)}[\|D_b(T_{a\to b}(E_{b\to a}(x_b))) - x_b\|_1].\end{aligned} \tag{1}$$

The corresponding data path through the network is shown in Figure 4(a). This may be thought of as an autoencoder loss, where the autoencoder, in addition to an encoder and a decoder, has a bipartite latent space, with a translator between its two parts.

Note that some previous unpaired translation works [26, 28], use an ordinary identity loss (without cross-translation), where images from domain B are fed to the $A \to B$ generator, and vice versa. We adopt this loss as well, as we found it to complement our cross-identity loss in Eq. (1):

$$\begin{aligned}\mathcal{L}_{Id} = \mathbb{E}_{x_a\sim p_{data}(A)}[\|D_a(E_{b\to a}(x_a)) - x_a\|_1]+\\ \mathbb{E}_{x_b\sim p_{data}(B)}[\|D_b(E_{a\to b}(x_b)) - x_b\|_1].\end{aligned} \tag{2}$$

3.3. Latent Cross-Translation Consistency

While the normal expected input, for each encoder, is an image from its intended source domain, let us consider the scenario where one of the encoders is given an image from its target domain, instead. For example, if A are images of horses and B images of zebras, what should happen when a zebra image x_b is given as input to the "horse-to-zebra" encoder $E_{a\to b}$? Our intuition tells us that in such a case we'd like the generator to avoid modifying its input. This implies that the resulting latent code $z_b = E_{a\to b}(x_b)$ should capture and retain the essential "zebra-specific" information present in the input image. The translator $T_{b\to a}$ is trained to map such "zebra features" to "horse features", thus we expect $\hat{z}_a = T_{b\to a}(z_b)$ to be similar to the latent code $z_a = E_{b\to a}(x_b)$, obtained by feeding the zebra image to the "zebra-to-horse" encoder, which should also yield "horse features". The above reasoning, applied in both directions, is formally expressed using the *latent cross-translation consistency loss* (see Figure 4(b)):

$$\begin{aligned}\mathcal{L}_{CTC} = \mathbb{E}_{x_a\sim p_{data}(A)}[\|T_{a\to b}(E_{b\to a}(x_a)) - E_{a\to b}(x_a)\|_1]+\\ \mathbb{E}_{x_b\sim p_{data}(B)}[\|T_{b\to a}(E_{a\to b}(x_b)) - E_{b\to a}(x_b)\|_1].\end{aligned} \tag{3}$$

3.4. Latent Cycle-Consistency

Our final latent space regularization is designed to ensure that our cross-translators are bijections between the two latent spaces of the generators. Similarly to the motivation behind the cycle-consistency loss of Zhu et al. [28], having bijections helps achieving a meaningful mapping between the two domains, as well as avoids mode collapse during the optimization process.

Specifically, we require that translating a latent code z_b first by $T_{b\to a}$ and then back by $T_{a\to b}$ yields roughly the same code back:

$$\begin{aligned}\mathcal{L}_{ZCyc} = \mathbb{E}_{x_a\sim p_{data}(A)}[\|T_{a\to b}(T_{b\to a}(z_b)) - z_b\|_1]+\\ \mathbb{E}_{x_b\sim p_{data}(B)}[\|T_{b\to a}(T_{a\to b}(z_a)) - z_a\|_1].\end{aligned} \tag{4}$$

The data path corresponding to this *latent cycle-consistency loss* is depicted in Figure 4(c).

3.5. Implementation and Training

Our final loss, which we optimize throughout the entire training process is a weighed sum of the losses presented in the previous sections, and a GAN loss [4],

$$\begin{aligned}\mathcal{L} = \lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{Id}\mathcal{L}_{Id}+\\ \lambda_{CTC}\mathcal{L}_{CTC} + \lambda_{ZId}\mathcal{L}_{ZId} + \lambda_{ZCyc}\mathcal{L}_{ZCyc}.\end{aligned} \tag{5}$$

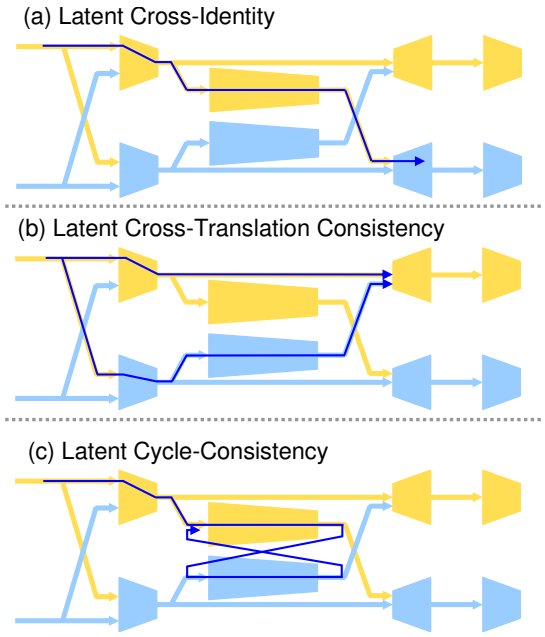

Figure 4: Data paths used by the novel loss terms when training our model (symmetric paths are omitted for clarity). (a) The latent cross-identity loss trains the cross-translators to map between the two latent spaces of the dual generators. (b) The latent cross-translation consistency loss regularizes the latent spaces generated by each of the two encoders. (c) The latent cycle-consistency loss ensures that the cross-translators define bijections between the two latent spaces.

Rather than using a negative log likelihood objective in the GAN loss, we make use of a least-squares loss [17]. Additionally, we adopt Shrivastava et al.'s method [21], which updates the discriminators using a history of translated images rather than only the recently translated ones.

Unless otherwise mentioned, in all of the experiments in this paper, we set $\lambda_{GAN} = 1$, $\lambda_{Id} = 3$, $\lambda_{CTC} = 3$, $\lambda_{ZId} = 6$ and $\lambda_{ZCyc} = 6$. We have tuned these hyperparameters by running a grid search, attempting to maximize the FID of CrossNet on a Style Transfer task.

For our generators, we adopt the Encoder and Decoder architecture from Johnson et al. [10]. Their Encoder architecture consists of an initial 7x7 convolution with stride 1, two stride 2 convolutions with a 3x3 kernel, and 9 residual blocks with 3x3 convolutions. The Decoder consists of two transposed convolutions with stride 2 and a kernel size of 3x3, followed by a final convolution with kernel size of 7x7 and hyperbolic tangent activations for normalization of the output range. For our discriminator networks, we use 70x70 PatchGANs [14], whose task is to classify whether the overlapping patches are fake or real.

Finally, our two latent code translators consist of 9 residual blocks that use 3x3 convolutions with stride 1.

For all of the applications which we present in the following sections, we trained our proposed method using two sets of ˜1000 images (a total of ˜2000 images). The final generator which we used for producing the results was selected as the result after training for 200 epochs.

4. Comparisons

4.1. Ablation study

For evaluating the competence and effect of our newly devised cross-consistency losses, we conduct an ablation study. In Figure 5 we show some results for the Horse $\leftrightarrow$ Zebra translation, demonstrating the visual effect of adding each one of our three latent space losses. Additionally, we compare our results to those of CycleGAN [28]. In all of these results, both $\mathcal{L}_{GAN}$ and $\mathcal{L}_{Id}$ were included in our training. We use the Frechet Inception Distance (FID) [5] as an objective quantitative measure of result quality. We calculate the FID between a synthesized set and a set of real images from the target domain.

As may be seen in Figure 5, through gradually adding the three cross-consistency losses, results improve. The best results are obtained when all three losses are included, as shown in the 6th column. The 7th (rightmost) column shows translation results generated using CycleGAN, where it may be seen that they are less successful: some zebra stripes remain when translating zebras to horses (3rd row), and not all horses are translated to zebras (rows 4,6). The FIDs decrease as we add more of our loss terms, reaching our best result once all 3 losses conjoin. It may be seen that the FIDs of our results are better that those of CycleGAN.

4.2. Unpaired image-to-image translation

In Figure 6 we show a variety of our unpaired image-to-image translation results (CrossNet), compared with CycleGAN. From top to bottom, we show image-to-image translation results for: Apples $\leftrightarrow$ Oranges and Summer $\leftrightarrow$ Winter images. All of our results were achieved with the full loss in (5), with the relative weights reported in Section 3.5. Qualitatively, it may be observed that in all three image-to-image translation tasks, CrossNet outperforms CycleGAN, providing better texture transfer, color reproduction, and also better structure (visible in the Apples $\leftrightarrow$ Oranges translations). Additionally, the FID shows superior results for CrossNet. Note that for producing the CycleGAN results, where possible, we used the existing pretrained models, made available by Zhu et al. [28].

5. Applications

In all three of the applications which follow, we align with contemporary work [1, 20, 24] in terms of the baselines (MUNIT, DRIT and CycleGAN) and the widely accepted measures for these comparisons (FID).

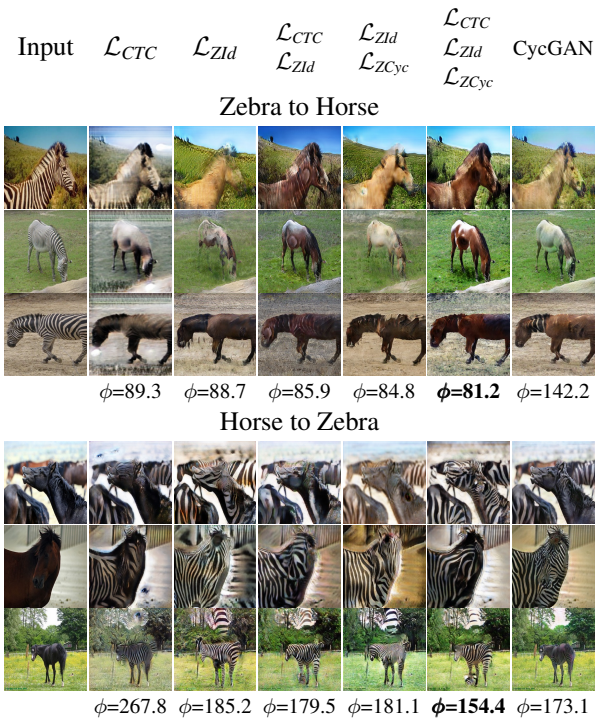

Figure 5: Ablation study: Through a gradual inclusion of losses, we demonstrate how the results of various translation tasks improve. A combination of all three cross-consistency losses (6th column) is shown to yield better results than those produced by CycleGAN (7th column) for the Horse $\leftrightarrow$ Zebra translations.

5.1. Specular $\leftrightarrow$ Diffuse

Most multi-view 3D reconstruction algorithms, assume that object appearance is predominantly diffuse. However, in real world images, often the contrary is true. In order to alleviate this restriction, Wu et al. [25] propose a neural network for transferring multiple views of objects with specular reflection into diffuse ones. By introducing a task specific loss, exploiting the multiple views of a single specular object, they are able to synthesize faithful diffuse appearances of an object. We make use of their publicly available dataset, and train CrossNet to translate between specular and diffuse objects.

Please note that differently from Wu et al., we do not rely on any prior assumptions specific to the task of specular to diffuse translation. A fundamental difference between Wu et al. and our approach is that the input to their network is an image sequence, which encourages the learning of a specific object's structure, while in our approach, only a single image is provided as input to the appropriate generator.

In Figure 7 we show the results of applying CrossNet on Wu et al.'s dataset. Results are shown for both translation directions, specular to diffuse and diffuse to specular. Both types of translations produce visually convincing results. It may be seen that our translation captures the fine details of the sculptures, and produces proper shading. An unintended phenomenon, are the changes which are created in the background. A perfect translation between these two domains shouldn't have altered any background pixels, but our method provides no means of controlling which pixels are left untouched. Additional results are provided in the supplementary material.

5.2. Mobile phone to SLR

Although extremely popular, contemporary mobile phones are still very far from being able to produce results of quality comparable to those of a professional DSLR camera. This is mostly due to the limitations on sensor and aperture size. In a recent work, Ignatov et al. [8] propose a supervised method, and apply it on their own manually collected large-scale dataset, consisting of real photos captured by three different phones and one high-end reflex camera. Here, we demonstrate our approach's applicability on the same task.

Ignatov et al.'s approach is fully supervised, where pairing of images (mobile and DSLR photos) is available during training. To achieve this, they introduced a large-scale DPED dataset that consists of photos taken synchronously in the wild by a smartphone and a DSLR camera. In contrast to DPED, we naively train CrossNet to translate iPhone photos to those of DSLR quality, ignoring the pairing and adding no additional explicit priors on the data.

Since the DLSR camera did not capture the scene from exactly the same perspective as the iPhone, a quantitative comparison is not possible. Ignatov at al. proposed aligning and warping between the two images, but since the entire essence of the approaches we compare is to produce high resolution details, we find warping (and thus resampling) improper and resort to a qualitative comparison. However, we report the FID as an objective method of comparison.

We compare our results to CycleGAN, MUNIT [7], DRIT [13] and to the fully supervised DPED method, and show that our approach generates translations that are richer in colors and details. The FID results show that DPED's supervised method yields the best results, but that among the unsupervised methods, our results ranks first.

In Figure 8 we present our results and compare them with those mentioned above. Zooming into specific regions of the results, shows the effectiveness of CrossNet vs. the competitors, for producing colors and details in the translated images. In order to properly compare, all of the algorithms were fed with an input image scaled to 256x256. We show the input, CrossNet, CycleGAN's, MUNIT's, DRIT's and DPED's results, from left to right respectively. The odd rows show the full images, while the even rows show a zoom-in. The two top rows show CrossNet's ability to en-

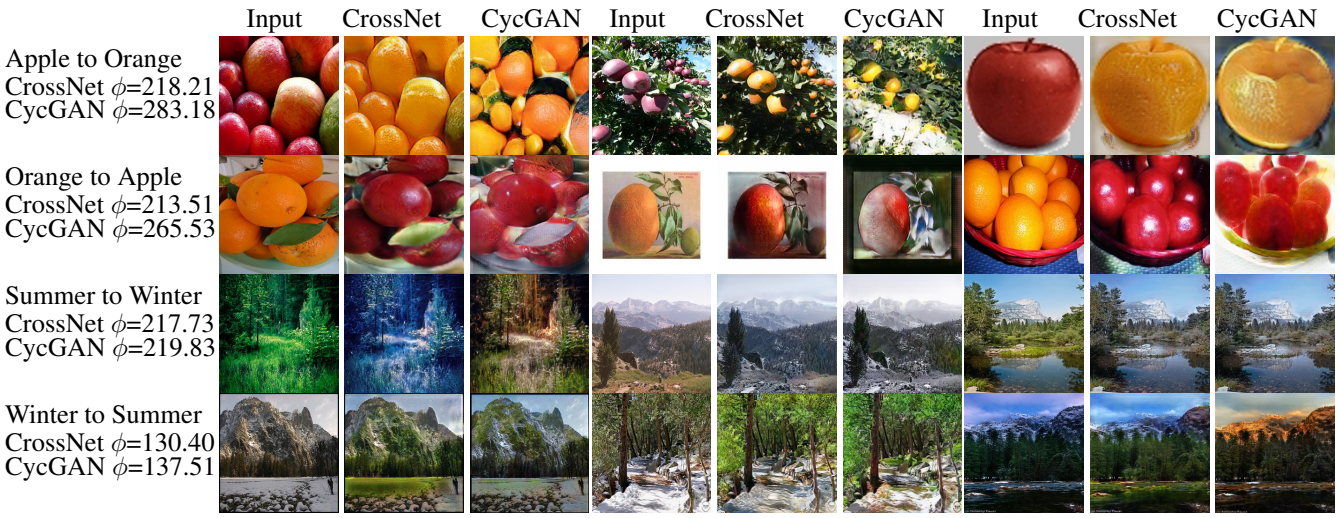

Figure 6: A variety of image-to-image translation results of our method (CrossNet), compared to CycleGAN. In the Apples $\rightarrow$ Oranges translations CycleGAN consistently produces errors and artifacts that seem semantic in nature (shadows become illuminated, the rightmost orange looks like a superimposed image of an open orange).

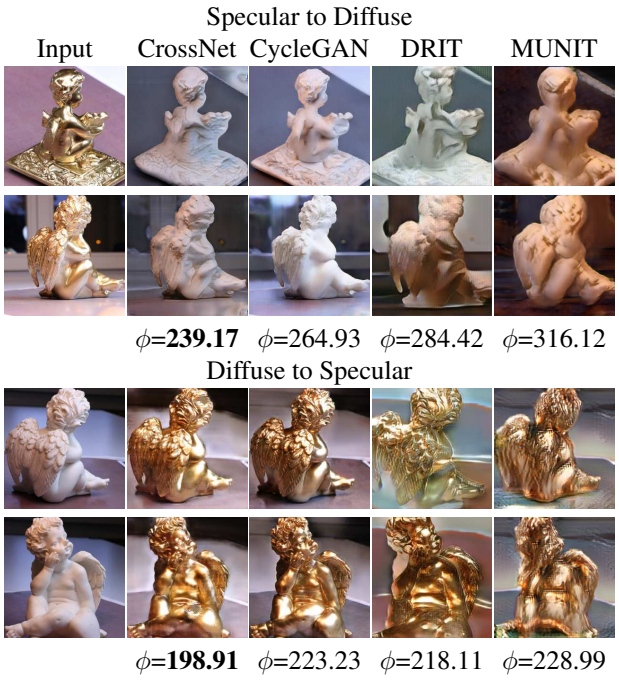

Figure 7: Specular $\leftrightarrow$ Diffuse translation: Columns show the inputs and various competing results. The two top rows show a translation from a specular input to a diffuse output, while the two bottom rows show the opposite translation.

hance colors better than the competing methods. The two middle rows show that our approach enhances contrast more strongly. Finally, the two bottom rows show that our result enhances details better than the competitors. We provide additional results in the supplementary material.

Figure 8: Results of mobile phone photo enhancement. We show the input, CrossNet, CycleGAN's, MUNIT's, DRIT's and DPED's results, from left to right. Odd rows show full images, while even ones show a zoom-in.

5.3. Semantic foreground extraction

Here, we demonstrate that using CrossNet it is possible to learn to extract a common object of interest from an image, by training on an unstructured image collection obtained using internet search engines. Specifically, we search Google images, once by using the search term "elephant", and once by using "elephant white background" to obtain two sets of images, with roughly 500 images in each set. CrossNet is then trained to translate between these two sets in an unpaired fashion. Note that this is a highly asymmetric translation task: while replacing all non-elephant pixels in an image with white color is well-defined, the opposite direction has a multitude of plausible outcomes. Thus, one can hardly expect a bijection between the original image domains, underscoring the importance of constraining and regularizing the latent spaces.

To obtain clearly extracted objects, we employ a slightly modified pipeline. Rather than using the raw outcome of the translation from images with a general background to ones with a white background, we use the result to define a segmentation mask by simple thresholding. Specifically, we replace near-white pixels with white color, while the color of the remaining pixels is replaced with their original values (before the translation). A pixel is considered nearly white if its luminance exceeds 243.

Figure 9 presents some of our results, applied on test images of elephants. We compare our results with CycleGAN, which were applied with the same post processing operation as our results. We also compare with AGGAN [18], which extracts attention maps which have been shown to be useful for segmentation. Additionally, we also compare our method with GrabCut [19]. Since GrabCut requires a initial bounding box, we provide it manually for the test images.

The results in Figure 9 show that although our method assumes no priors on the segmentation task at hand, it effectively extracts the foreground objects. We attribute this success to the presence of all of the necessary information in the latent code, making our approach more adequate for this task. AGGAN is shown to fail on this task, most probably due to the same class of images found in the two input sets, confusing its attention mechanism. Note how its attention masks do not indicate any clear attention region. Additional results are provided in the supplementary material.

In addition to our manually composed dataset of elephants, we quantify our method's results by employing the UT Zappos50K shoes dataset [27] which present photos with white backgrounds as well as on the HDSeg dataset [12] of Horses with ground truth segmentation masks. During training to segment shoes, we synthetically replaced the white background with natural scenery photos.

We report the RoC curves of CrossNet, CycleGAN and AGGAN in Figure 10, as well as the Area Under Curve (AUC) through which CrossNet's superiority is noticeable.

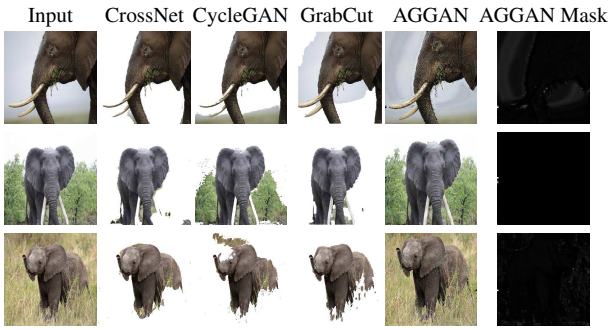

Figure 9: Results of employing CrossNet for foreground extraction, applied on elephant photos. From left to right, we show the input image, CrossNet, CycleGAN, GrabCut and AGGAN.

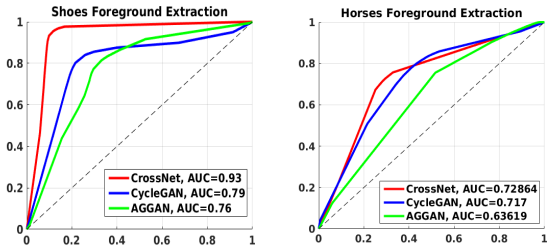

Figure 10: Objective evaluations comparing between CrossNet, CycleGAN and AGGAN for foreground extraction on the UT Zappos50K dataset and HDSeg.

6. Discussion and Future work

We have presented CrossNet, an architecture for generic unpaired image-to-image translation. During training, CrossNet imposes constraints on the latent codes produced by the generators, rather than imposing them in the image domains. Latent codes constitute a compact representation of the aspects of the data that are most relevant to the task that the model is trained for. Since the unpaired translation setting is challenging and under-constrained, proper regularization of these latent spaces is crucial, and we demonstrate that models trained with such regularizers are more effective at their intended task.

An interesting observation is that all our losses are symmetric. They operate in both directions, A to B and B to A. Although we have shown pleasing results when applying our method to the proposed applications, in some of them, the asymmetric nature of the task suggests that asymmetric constraints might be more appropriate. For example, in the foreground extraction application, background removal is well defined, but adding a background is not, having a huge space of plausible outcomes. The exploration of asymmetric architectures and constraints is another promising direction for future work.

References

[1] Y. Alharbi, N. Smith, and P. Wonka. Latent filter scaling for multimodal unsupervised image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1458–1466, 2019.

[2] 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. *arXiv preprint*, 1711, 2017.

[3] H. Dong, P. Neekhara, C. Wu, and Y. Guo. Unsupervised image-to-image translation with generative adversarial networks. *arXiv preprint arXiv:1701.02676*, 2017.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[5] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[6] Y. Hoshen and L. Wolf. Nam: Non-adversarial unsupervised domain mapping. *arXiv preprint arXiv:1806.00804*, 2018.

[7] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. *arXiv preprint arXiv:1804.04732*, 2018.

[8] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *the IEEE Int. Conf. on Computer Vision (ICCV)*, 2017.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[10] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[11] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[12] A. Kolesnikov, M. Guillaumin, V. Ferrari, and C. H. Lampert. Closed-form approximate crf training for scalable image segmentation. In *European Conference on Computer Vision*, pages 550–565. Springer, 2014.

[13] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.

[14] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[15] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[16] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[17] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017.

[18] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3697–3707, 2018.

[19] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[20] Z. Shen, M. Huang, J. Shi, X. Xue, and T. Huang. Towards instance-level image-to-image translation. *arXiv preprint arXiv:1905.01744*, 2019.

[21] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, volume 2, page 5, 2017.

[22] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017.

[23] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. 2017.

[24] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5849–5859, 2019.

[25] S. Wu, H. Huang, T. Portenier, M. Sela, D. Cohen-Or, R. Kimmel, and M. Zwicker. Specular-to-diffuse translation for multi-view reconstruction. In *European Conference on Computer Vision*, pages 193–211. Springer, 2018.

[26] Z. Yi, H. R. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017.

[27] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.