9-2020

Weakly paired multi-domain image translation

M.Y. ZHANG

Zhiwu HUANG
Singapore Management University, zwhuang@smu.edu.sg

D.P. PAUDEL

J. THOMA

Gool L. VAN

Weakly Paired Multi-Domain Image Translation

Marc Yanlong Zhang[1]
zhangma@student.ethz.ch
Zhiwu Huang[1]
zhiwu.huang@vision.ee.ethz.ch
Danda Pani Paudel[1]
paudel@vision.ee.ethz.ch
Janine Thoma[1]
jthoma@vision.ee.ethz.ch
Luc Van Gool[12]
vangool@vision.ee.ethz.ch

[1] Computer Vision Lab
ETH Zurich
Switzerland
[2] ESAT-PSI
KU Leuven
Belgium

Abstract

In this paper, we aim at studying the new problem of weakly paired multi-domain image translation. To this end, we collect a dataset that contains weakly paired images from multiple domains. Two images are considered to be weakly paired if they are captured from nearby locations and share an overlapping field of view. These images are possibly captured by two asynchronous cameras—often resulting in images from separate domains, e.g. summer and winter. Major motivations for using weakly paired images are: (i) performance improvement towards that of paired data; (ii) cheap labels and abundant data availability. For the first time in this paper, we propose a multi-domain image translation method specifically designed for weakly paired data. The proposed method consists of an attention-based generator and a two-stream discriminator that deals with misalignment between source and target images. Our method generates images in the target domain while preserving source image content, including foreground objects such as cars and pedestrians. Our extensive experiments demonstrate the superiority of the proposed method in comparison to the state-of-the-art. The new dataset and the source code are available at https://github.com/zhangma123/weaklypaired.

1 Introduction

When dealing with image-to-image translation, most of the state-of-the-art deep learning models are either fully supervised or unsupervised. Fully supervised training is very efficient since the generated image can be directly compared to the desired outcome, i.e. a target image. Unsupervised models and losses, on the other hand, do not require strongly paired data. The advantage is apparent: Strongly paired data is usually difficult to acquire. A major drawback of unsupervised methods—in addition to the training instability—is the need of additional hard constraints like cyclic consistency [29], which often results in large networks that are difficult to train. This raises the question of *whether weakly paired data*

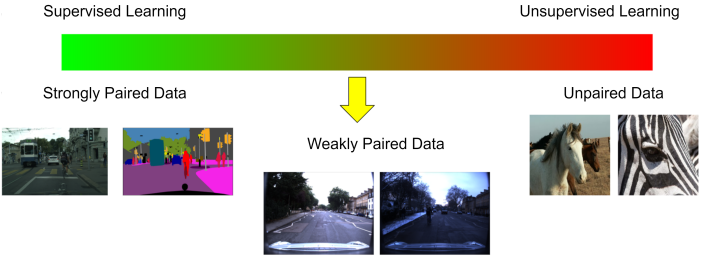

Figure 1: Examples of paired, weakly paired, and unpaired data used for image translation.

can be efficiently used to train deep learning models? We consider two images to be weakly paired if they are captured from nearby locations and share an overlapping field of view (see Figure 1). These images may be captured by two cameras at very different instants of time—often resulting in images from separate domains, e.g. summer, winter, day, and night.

In essence, weakly paired images are neither pixel-wise aligned (as in the strongly paired case) nor do they lack pixel-wise correspondences (as in the unpaired case). We argue that the acquisition of weakly paired images, unlike that of strongly paired ones, is easy and the current availability is abundant. Our requirement for "weakly pairedness" is that the geographic location and orientation should be roughly the same, sharing some overlapping field of view. This setup does not put any restrictions on foreground objects and image style. The requirements for weakly paired images can easily be satisfied by using independently moving cameras which roughly look towards the same scene, or by using sensory data such as GPS-tags and compass information. In this work, we highlight the fact that setups using weakly paired images have gathered little to no attention in the literature. To address this, this paper proposes a novel method for the first time, to the best of our knowledge. Two major motivations of using weakly paired images are: (i) performance improvement towards that of paired data; (ii) cheap labels and abundance in data availability.

While being a promising alternative to the paired case, weakly paired data comes with its own share of challenges, which include misalignments and the change in dynamic scene parts. For example, in the data used in this paper (i.e. the Oxford Robotcar [14]), images are taken by a camera mounted on a car. As the car travels along a fixed route during different seasons and weather conditions, weakly paired images from different domains are gathered together with the location and orientation of the vehicle. Since these images are inevitably misaligned and differ in terms of dynamic content (due to inaccurate GPS data, change in traffic/construction, or other scene dynamics), a method devised to handle weakly paired data must avoid the direct pixel-to-pixel comparisons. In this work, we make multiple necessary considerations to handle such data and demonstrated the effectiveness of the developed method for the intended task of image translation, as well as one of its applications for image retrieval. The major contributions of this paper can be summarized as follows:

- We formulate and address the problem of multi-domain image translation leveraging weakly paired images. To facilitate research in direction, we curated a the Oxford Robotcar Dataset [14] into a large scale multi-domain weakly paired images.
- We design a novel method that uses an attention model to tackle dynamic parts during image generation, a joint distribution learning concept for exploiting weakly paired data, and image classification for multi-domain image translation.

- In addition to the improvements in image translation, we provide an experimental evaluation demonstrating the superiority of the proposed method for image retrieval.

2 Related Works

The task of image-to-image translation can be broadly classified into two categories: paired and unpaired. The paired translation methods assume that the pixel-wise mapping between source and target is known [4, 9, 25]. On the other hand, unpaired image translation methods [12, 27, 29], use sets of images, each set representing a certain style or domain. Some extensions of these methods are used for many applications. For example, [2] uses translation model to improve the retrieval accuracy of night images. Similarly, [19] uses the same for foggy image segmentation. A key extension includes image translation in multi-domain scenarios, where images are translated to the desired target domain using deep networks [1, 8, 10, 11, 15, 20, 22, 26]. In this regard, three notable works include, StarGAN [5], GANimation [17], and SPADE [16].

StarGAN [5]. StarGAN alleviated the need for $n(n-1)$ networks for n domains by encoding the target domain information into the generator input. Using cycle consistency, StarGAN generates images on multiple target domains without requiring strongly paired data. Similar to CycleGAN [29], this cycle consistency forces the network to preserve key information for reconstructing the original image. A class prediction loss is introduced to ensure that the generator produces images in the correct domain. Furthermore, [5] also introduced a method to train a single model on multiple datasets by applying a mask vector to the label condition. Following [5], **StarGANv2 [6]** suggests to replace the domain label with a domain-specific style code that is able to represent more diverse styles of a specific domain using a mapping network and a style encoder. The mapping network is trained to transform random noise into a style code, while the encoder learns to extract the style code from a given reference image.

GANimation [17]. In contrast to StarGAN's discrete nature of the domains's definition, GANimation synthesizes images on a continuous spectrum. GANimation is mainly tested for human facial features, such as emotions, that are best represented by continuous action units. GANimation also uses an attention network to mask areas requiring no modifications, whereas cycle consistency is used to preserves important information during translation.

SPADE [16]. Similar to pix2pixHD [9], SPADE generates photo-realistic conditioned upon a semantic image. The semantic images are used for spatially variant denormalization, which allows the realistic texture synthesis for a uniform semantic patch. On the generator side, a gradual upsampling scheme is on the input segmentation map to feed semantics in to SPADE blocks up to the desired output image size. The discriminator uses a multi-scale architecture which learns the joint distribution between semantic and RGB images.

3 Problem Formulation

Let us define the set of N different image domains as $\{\mathcal{X}_i\}_{i=1}^{N}$. For each pair of domains $\{\mathcal{X}_i, \mathcal{X}_j\}$, we use pairs of weakly-aligned images $\mathbf{P}_{ij}^k = \{\mathbf{I}_i^k \in \mathcal{X}_i, \mathbf{I}_j^k \in \mathcal{X}_j\}$, which are captured from nearby locations and share an overlapping field of view. The task of weakly paired multi-domain image translation aims at learning a mapping function $\psi_\theta(i) : \mathbf{I} \rightarrow \hat{\mathbf{I}} \in \mathcal{X}_i$, which maps any image $\mathbf{I}$ to domain $\mathcal{X}_i$, given the target domain. In the context of this paper, ψ is a convolutional neural network and θ are the network parameters. For the task

WNO	WNR	WDSn	WDO-1	WDO-2	WDSu	WDO-3	SDS-1	SDO	SDS-2	SDR
Winter	Winter	Winter	Winter	Winter	Winter	Winter	Summer	Summer	Summer	Summer
Night	Night	Day	Day	Day	Day	Day	Day	Day	Day	Day
Overcast	Rain	Snow	Overcast	Overcast	Sun	Overcast	Sun	Overcast	Sun	Rain

Table 1: Summary of the obtained data with corresponding attributes used. The domain is abbreviated and shown the top-most row. These shorthands are used throughout the paper.

of multi-domain image translation, we wish to learn θ using a given set of weakly paired images $\{\mathbf{P}_{ij}^k\}$. Inspired by [5], we aim at learning a single multi-domain model that translates any image into the target domain, by avoiding $O(N^2)$ complexity of training N^2 separate cross-domain models. Note that the pairs $\mathbf{P}_{ij}^k$ are neither unpaired nor paired, as commonly assumed for the task of image translation. To leverage such weakly paired data, it becomes necessary to design a new model that is different from the cases of paired or unpaired data.

4 Data Collection

We use the Oxford Robotcar Dataset [14] to train and test our model, after carefully curating the original dataset making it suitable for the problem at hand. It consists of video sequences taken from a driving car acquired over a period of a year along a fixed driving route. These video frames come with INS (Inertial Navigation System) data, which is used to pair the images. The pairing takes place between several different weather, season, and lightning conditions. Inherently, these images are *weakly paired* since dynamic objects will inevitably change. Furthermore, two images in a pair are very likely to be misaligned due to imperfect INS data and possibly other traffic conditions. We pair the images (frames extracted from videos) such that the geometric distance between two images in a pair is within a chosen threshold. Using this method, we were able to collect more than 120k images. A summary of the curated dataset is provided in Table 1.

5 Proposed Approach

In the suggested scenario of weakly paired image translation, any paired images are captured from nearby locations, but are generally misaligned and contain varied dynamic content. In general, this disables the use of those methods with direct pixel-to-pixel supervision. To address this, our solution is the exploitation of generative distribution learning, which typically enforces a generator to approximate the distributions of images from target domains so that the input can be transformed to the given target domains. In particular, the key idea of our approach is to learn multi-domain mappings by exploiting the technology of generative adversarial networks (GANs) [7] that has shown its strong capability of distribution learning. To this end, we design an attention-based generator to handle dynamic parts when learning the mapping set between multiple domains, while exploiting a two-stream discriminator to align the feature maps of misaligned inputs for better marginal and joint distribution learning on the weakly paired multi-domain real and produced images. The generator and the discriminator are jointly optimized by a min-max objective function. The overview of the proposed model architecture is illustrated in Figure 2.

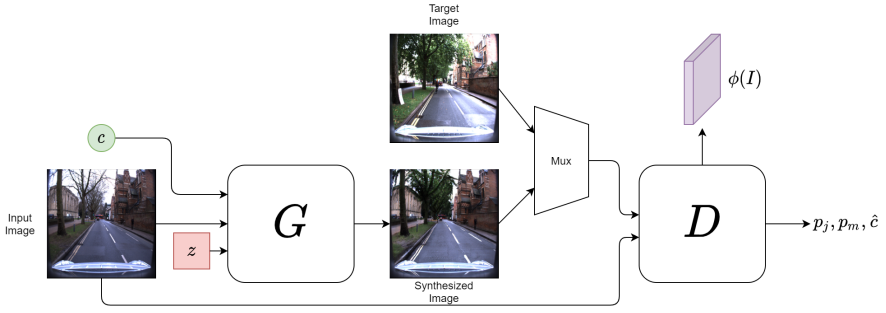

Figure 2: **Overall** architecture. The generator takes an input image, a class label c and prior z and synthesizes an image. "Mux" refers to a multiplexer that chooses between the real and fake images during training. The discriminator takes either an input+fake pair or an input+real pair and outputs the joint, marginal, and class predictions p_j, p_m and $\hat{c}$, respectively. The intermediate features $\phi(I)$ extracted from the discriminator are used to compute $\mathcal{L}_{feat}$.

5.1 Network Architecture

Generator. The structure of our generator is based on the SPADE generator [16]—given its superiority for paired image translation—where we change the output layers to produce a color mask $\mathbf{C}$ as well as an attention mask $\mathbf{M}$. Following the attention-based residual learning of GANimation [17], the generated image is obtained by,

$$\mathbf{I}_{Out} = \mathbf{M} * \mathbf{I}_{In} + (\mathbf{1} - \mathbf{M}) * \mathbf{C}. \tag{1}$$

The color mask $\mathbf{C} \in \mathbb{R}^{3\times H\times W}$ describes the color information that the generator paints on top of the input while the attention mask $\mathbf{M} \in [0,1]^{1\times H\times W}$ describes the position where the color painting is performed. This architecture enables the model to preserve background while changing the the foreground objects as well as the style of the scene. We noticed significant qualitative improvements when using an attention-based approach over the original generator structure of [16]. For conditioning on the desired target domain, we use a binary encoding of the label and concatenate it with a prior z which is sampled from a normal distribution. Additionally, we introduce $z_{align} \in \mathbb{R}^{1\times H\times W}$ and concatenate it to the input image along the channel dimension. This *alignment noise* allows the generator to synthesize differently aligned images, which is also sampled from a normal distribution (independent of the prior z) during training and inference.

Discriminator. We employ a multi-scale discriminator architecture similar to that of SPADE, i.e. real/fake prediction is performed on two different image scales. Similarly, intermediate layer activations are extracted to compute the GAN Feature Loss. For marginal as well as joint distribution learning, a marginal discriminator is added in parallel to the original joint discriminator. The marginal discriminator also outputs an additional class label prediction. With such generator, it is very likely that the generated image is more aligned to the input than to the target. This allows the discriminator to distinguish between fake image pairs (input + fake) and real image pairs (input + real) based on their alignment to the input, resulting in a too powerful discriminator that fails to properly guide the training process of the generator. To alleviate this issue, we divide the discriminator input paths into two separate streams, D_1 and D_2. The underlying motivation is to use these intermediate layers to transform images into a representation where alignment is not a dominant factor.

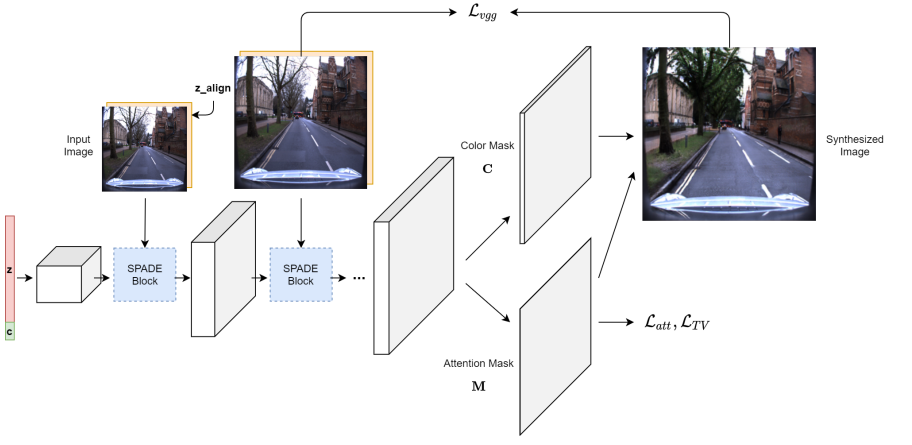

Figure 3: **Generator** architecture. To the left, the prior z (red) is sampled from a normal distribution and concatenated with the domain label (green). The alignment noise *z_align* is shown in yellow behind the input image. The input image is fed into the generator at different scales via SPADE Blocks (blue) which are directly taken from SPADE [16]. The input image together with the Color and Mask are fused into the synthesized image. Mask regularizations $\mathcal{L}_{att}$ and $\mathcal{L}_{TV}$ ensure that the mask is smooth and well saturated. The perceptual loss $\mathcal{L}_{vgg}$ is computed between the input and synthesized image using a pretrained VGG19 network.

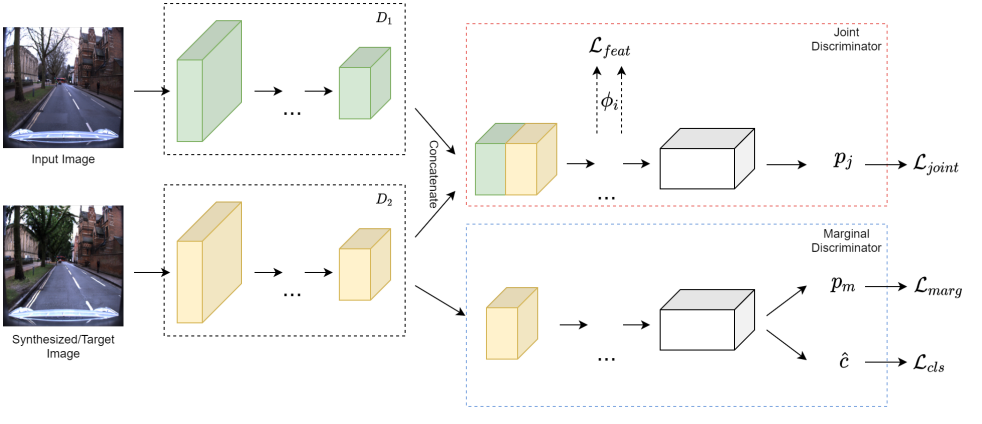

Figure 4: **Discriminator** architecture. D_1 and D_2 are two networks that handle input and fake/real image respectively. Their outputs are concatenated and further processed by the joint discriminator where features ϕ_i are extracted from the intermediate layers to compute $\mathcal{L}_{feat}$. Both the joint and marginal discriminator produce predictions p_j, p_m on whether the input was synthesized by the generator or sampled from the data to compute the losses $\mathcal{L}_{joint}$ and $\mathcal{L}_{marg}$ respectively. The marginal discriminator only has access of the output of D_2 which also outputs a class prediction $\hat{c}$. The prediction $\hat{c}$ is later used to compute the loss $\mathcal{L}_{cls}$.

5.2 Optimizing the Model

Adversarial Loss. We use joint and marginal adversarial losses to train our generator and discriminator, denoted by $\mathcal{L}_{joint}$ and $\mathcal{L}_{marg}$, respectively. This is to ensure that the generator synthesizes images that are indistinguishable from real images.
GAN Feature Loss. Following SPADE [16], we extract intermediate features ϕ_i of the joint discriminator and compute the L1 distance between the features of the synthesized and target images. This loss is denoted by $\mathcal{L}_{feat}$. Layer-wise loss terms are combined using positively weighted mean, with higher weights to lower-level features.
VGG Loss. This perceptual loss is computed by taking the L1 distance between features extracted by a pretrained VGG19 network. In contrast to $\mathcal{L}_{feat}$, this loss is computed between the synthesized and *input* image. We also weigh each layer by assigning higher weights to higher-level features in order to preserve as much semantic details as possible. Note that $\mathcal{L}_{vgg}$ and $\mathcal{L}_{feat}$ have opposite objectives: VGG Loss tries to pull the image closer to the input, while the GAN Feature Loss tries to push the image closer to the target.
Classification Loss. Similar to StarGAN [5] we also introduce a classification loss by comparing the class prediction $\hat{c}$ with the ground truth domain label. We use Binary Cross Entropy for this loss, on the predictions for synthesized as well as the target images.
Mask Regularization. We regularize the attention mask $\mathbf{M}$ by rewarding smooth masks (minimize the total variation of the mask $\mathcal{L}_{TV}$). In contrast to GANimation [17], we penalize "dark masks", i.e. we assign a higher loss if the mask is close to zero since that would make the mask redundant. The loss corresponding to this regularization is denoted by $\mathcal{L}_{att}$.
Full Objective. The full training objectives $\mathcal{L}_G$ and $\mathcal{L}_D$, respectively for the generator G and the discriminator D are derived from two sets of losses, $\mathcal{S}_G = \{\mathcal{L}_{joint}, \mathcal{L}_{marg}, \mathcal{L}_{feat}, \mathcal{L}_{vgg}, \mathcal{L}^f_{cls}, \mathcal{L}_{att}, \mathcal{L}_{TV}\}$ and $\mathcal{S}_D = \{\mathcal{L}_{joint}, \mathcal{L}_{marg}, \mathcal{L}^r_{cls}\}$. The $\mathcal{L}_G$ (resp. $\mathcal{L}_D$) combines the losses of $\mathcal{S}_G$ (resp. $\mathcal{S}_D$) using a set of appropriate hyperparameters. Please, refer to the supplementary material for the details about our hyperparameters. In terms of the loss computation complexity, despite of used several loss terms, our model is only marginally higher compared to SPADE [16]. More specifically, only the tail of the generator and the head of the discriminator are expanded, accordingly which result in significant improvement in performance for our problem. Moreover, majority of loss terms respect the same setup as in SPADE [16] and GANimation [17], making the task of hyperparameter tuning straightforward.

6 Experiments

6.1 Multi-Domain Image Translation

Evaluation Metric. Finding suitable metrics for the evaluation of our model is not straightforward. Similarity scores such as the Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR) are not suited for weakly paired data since the generated image will always deviate from the target image. Therefore, we use other quantities such as the Fréchet Inception Distance (FID), which compares image distributions. Other learned metrics such as LPIPS [28] are not meaningfully applicable: Since the generated image is directly compared with the target, large foreground objects significantly decrease the score (increase the distance). We therefore propose the **average SIFT descriptor distance**. This metric is motivated by the misalignments and changes in dynamic objects in each scene. One reasonable approach is to first find the Fundamental Matrix that relates keypoints between two images. Given

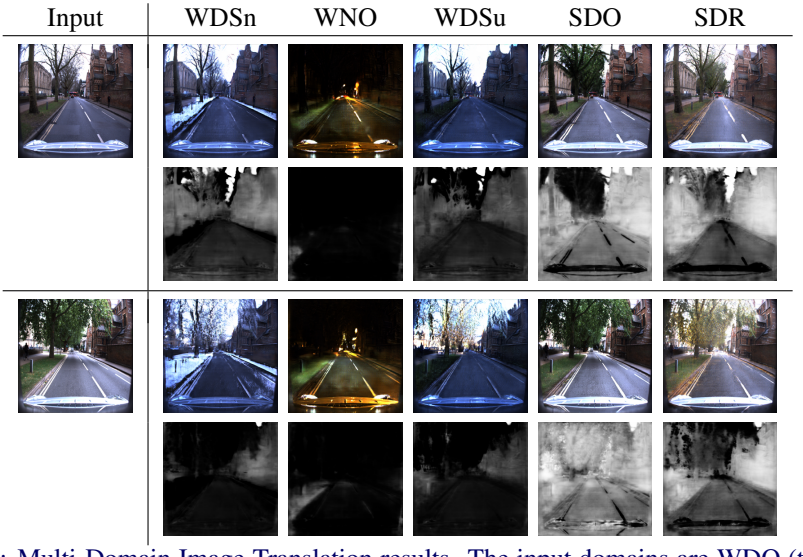

Figure 5: Multi-Domain Image Translation results. The input domains are WDO (top) and SDO (bottom). This figure shows that our model is capable of translating images into different target domains as indicated by the column description. The attention masks are depicted in the second and forth row, respectively.

two images, we detect and compute SIFT descriptors of both images and use a FLANN algorithm (KNN, k=2) to match these descriptors. Since there are many matches and most of them are ambiguous, we use a ratio test to filter these out as proposed by Lowe [13]. Afterward, RANSAC is used to calculate the Fundamental Matrix and the inliers. Using these keypoints we then can extract the SIFT descriptors and compute the average distance (L1 norm) between them. To accommodate for different numbers of matches we normalize the score by dividing by the total number of matches for each image pair as well as by 128.

Comparison to Baseline Models. We refer to [18] for an overview of state-of-the-art models that can achieve multi-domain image-to-image translation. The two most notable ones are **StarGAN** [5] as well as **GANimation** [17]. Despite both of them dealing with facial expressions, these models come closest to our task. The most recent work, StarGANv2 [6] improves many aspects of StarGAN and can also deal with animal faces. Unfortunately, we were unable to produce meaningful results using StarGAN and StarGANv2. We believe that this failure can be attributed to the following: 1) StarGAN and StarGANv2 were originally designed for the translation of faces rather than scenes which is much more challenging. 2) The one-way cyclic constraint of both models may be too weak to preserve meaningful input features. 3) StarGANv2 uses a style encoding network that produces latent codes for different domains which may not be sufficient to condition the generator output. We therefore only use GANimation as a baseline model, training the model in a fully unsupervised setting, using GANimation's default hyperparameters. As a baseline relying on strongly paired images, we use SPADE in its original form. For all the competing methods, we use about 30k training data. Figure 6 and Table 2 report qualitative and quantitative results. By comparison, our proposed method generally shows its superiority over competing models.

Ablation Study. We study the relative importance of the different components of our model: alignment noise, classification loss, attention-based modeling, and stacking the discriminator

Figure 6: Ablation Study and Model Comparison. We compare the final method to GANimation [17] and SPADE [16] as well as the ablation model that disregards the attention mask denoted by "no_mask". The right-most column shows the weakly paired target images.

Domain	no_gen_align		no_classify_loss		no_mask		stack_discr		SPADE_org		GANimation		FINAL	
	FID	SIFT	FID	SIFT	FID	SIFT	FID	SIFT	FID	SIFT	FID	SIFT	FID	SIFT
WNO	62.772	10.186	**60.642**	10.120	66.276	10.283	70.334	10.152	66.903	**10.056**	74.816	10.836	62.675	10.836
WNR	55.173	10.159	57.077	10.570	65.121	10.750	51.419	10.330	**46.783**	10.594	103.225	11.483	53.637	**10.151**
WDSn	63.477	11.258	74.535	11.579	70.966	11.735	**60.901**	11.584	68.767	11.719	84.556	11.691	77.133	**11.006**
WDO-1	54.649	8.244	52.958	8.397	58.393	8.617	53.573	8.158	**46.951**	**8.004**	75.065	9.644	54.092	8.253
WDO-2	53.724	8.732	52.938	8.837	57.033	9.017	52.704	8.716	**48.441**	**8.569**	74.516	9.840	53.806	8.685
WDSu	**64.514**	**10.622**	69.056	11.114	74.615	11.186	72.149	10.732	72.379	11.036	80.119	11.152	65.945	11.011
WDO-3	97.690	10.355	94.172	10.374	97.955	10.464	99.759	10.298	94.482	**10.159**	98.172	10.732	**89.235**	10.191
SDS-1	58.931	10.019	59.255	10.213	63.408	10.553	62.814	10.326	61.970	10.281	93.415	11.218	**54.354**	**9.949**
SDO	54.435	10.265	**52.428**	10.212	55.840	10.427	55.539	10.160	52.511	10.148	73.000	10.696	53.578	**10.096**
SDS-2	85.206	10.993	84.859	11.236	88.249	11.188	80.861	11.063	79.759	11.752	99.548	11.306	**79.426**	**10.471**
SDR	73.851	10.811	**61.618**	10.651	65.668	10.787	68.473	10.463	69.403	**10.249**	79.138	11.162	65.490	10.360
Mean:	65.857	10.150	65.413	10.300	69.411	10.455	66.230	10.180	**64.395**	10.233	85.052	10.887	64.488	**10.092**
Std:	14.262	0.912	13.877	0.958	13.176	0.916	14.533	0.971	15.072	1.144	**11.414**	**0.643**	12.504	0.886

Table 2: FID scores and average SIFT descriptor distances. Each column indicates the model that is used and is subdivided into FID Score (left) and SIFT descriptor distance (right). Lower is better in both cases. The lowest FID and SIFT values per row are marked in **bold**.

input. The last aspect involves concatenating the (input + real/fake) pair at the input of the discriminator instead of forwarding the images through two separate networks, namely D_1 and D_2. Figure 6 shows some results of the ablation study. Only the results of "no_mask" are depicted, i.e. if we regress the output directly without an attention mechanism, although other ablation models were considered as well which are listed in Table 2 along with their scores. Looking at these results we can conclude the following: The motivation behind the *alignment noise* was to allow the generator to output differently aligned images. However, the testing results show that this is not the case, meaning that the noise was ignored. It may act as regularization which improves the model's overall generalization capability, therefore increasing the score. When disabling the *classification loss* we only noticed a slight decrease in image quality. Unsurprisingly, removing the *mask* and falling back to the original SPADE generator architecture drastically decreases the image quality, both quantitatively and qualitatively. The model's ability to directly apply details from the input is a substantial advantage. Lastly, concatenating the images at the beginning of the discriminator substantially decreases the quality, as we had hoped.

6.2 Multi-Domain Image Retrieval

One application that benefits from image domain translation is multi-domain image retrieval. In a setting where query images are taken under different conditions than the reference images, retrieval can be facilitated by translating all images into the same domain prior to

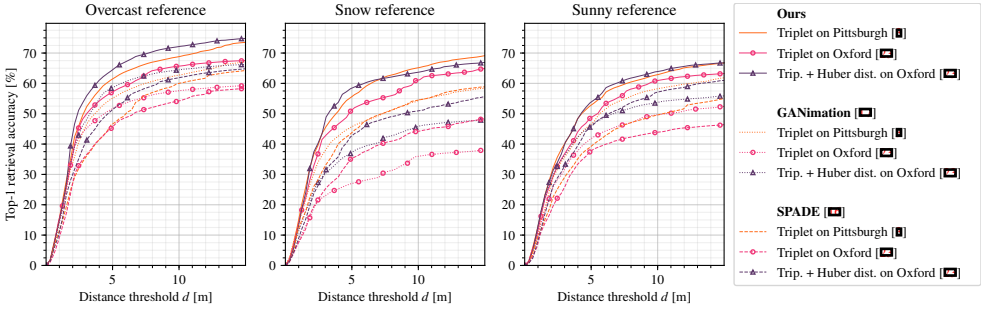

Figure 7: Top-1 retrieval accuracy on the Oxford Robotcar Dataset for three different reference conditions as a function of the distance threshold d for correct retrieval. The plots highlight the influence of image domain translation as an initial step for easier cross-domain image retrieval. We compare our method to GANimation [17] and SPADE [16] by combining each with three different image retrieval networks from [8] and [23].

retrieval. In Figure 7 we compare the image retrieval accuracy obtained with the help of our translation to that of GANimation [17] and SPADE [16]. For the retrieval step, we extract global image features using a VGG-16 [21] network followed by NetVLAD [8] spatial pooling. This network is initialized with weights trained on the Pittsburg Dataset [24] as described by [8] and weights trained on the Oxford RobotCar Dataset [14] as described by [23]. For every query image, we then retrieve its matching reference image via nearest neighbor search in the feature space. An image is considered correctly retrieved if the true query image coordinates lie within a distance threshold d of the retrieved reference image. As testing data, we use a region of the Oxford Robotcar Dataset which is geographically disjoint from the training region of [23]. We report the retrieval results for three different reference conditions: overcast, snow, and sunny. For each reference, the other two conditions are used as queries. The total number of testing images is 24590. Figure 7 shows, our translation as an initial step for image retrieval yields consistently better retrieval accuracy.

7 Conclusion

In this paper, we have designed and trained a generative adversarial network based model for multi-domain image translation, which leverages weakly paired images. Although the problem of exploiting weakly paired data is new, we demonstrated—by processing a publicly available benchmark dataset—that the required weakly paired images are easy to obtain. To evaluate the translation performance under the proposed experimental setup, a new metric is introduced. Additionally, images translated using the proposed method were used for the task of image retrieval based localization. Our experimental results demonstrate both quantitative and qualitative benefits of using the weakly paired images for translation as well as the translation followed by retrieval. The proposed model does not require any cyclic architecture, therefore is superior in terms of the computational efficiency and its simplicity. We believe that this opens up a new direction for image translation and beyond, where cheaply acquired weakly paired data can be efficiently utilized to train deep networks.

References

[1] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 783–790, 2018.

[2] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019.

[3] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2015.

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[6] 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.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Le Hui, Xiang Li, Jiaxin Chen, Hongliang He, and Jian Yang. Unsupervised multi-domain image translation with domain-specific encoders/decoders. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2044–2049. IEEE, 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[10] Ye Lin, Keren Fu, Shenggui Ling, and Cheng Peng. An efficient multi-domain framework for image-to-image translation. *arXiv preprint arXiv:1911.12552*, 2019.

[11] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in neural information processing systems*, pages 2590–2599, 2018.

[12] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[13] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94. URL http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

[14] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36 (1):3–15, 2017. doi: 10.1177/0278364916679498. URL http://dx.doi.org/10.1177/0278364916679498.

[15] Xudong Mao and Qing Li. Unpaired multi-domain image generation via regularized conditional gans. *arXiv preprint arXiv:1805.02456*, 2018.

[16] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[17] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[18] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. Smit: Stochastic multi-label image-to-image translation, 2018.

[19] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7374–7383, 2019.

[20] Yangyun Shen, Runnan Huang, and Wenkai Huang. Gd-stargan: Multi-domain image-to-image translation in garment design. *PloS one*, 15(4):e0231719, 2020.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Hao Tang, Dan Xu, Wei Wang, Yan Yan, and Nicu Sebe. Dual generator generative adversarial networks for multi-domain image-to-image translation. In *Asian Conference on Computer Vision*, pages 3–21. Springer, 2018.

[23] Janine Thoma, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Geometrically mappable image features. *IEEE Robotics and Automation Letters*, 5(2):2062–2069, 2020.

[24] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890, 2013.

[25] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[26] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5914–5922, 2019.

[27] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.