Supplementary Material for "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

This supplementary material contains training details as well as more results for multi-domain image translation and retrieval.

1 Training Details

1.1 Detailed Training Objective

The loss function comprises of a collection of smaller learning objectives, each weighted according to their importance. In the following we enumerate the equations and provide a brief explanation for their choices.
The standard adversarial losses are given in (1) and (2). While $\mathcal{L}_{joint}$ tries to find the joint distribution, $\mathcal{L}_{marg}$ aims at finding the marginal distribution. Here, x and y are sampled from two different (arbitrary) domains and are therefore weakly paired. Note, that y represents an image from domain c. D_j and D_m denote the joint and marginal discriminator, respectively.

$$\mathcal{L}_{joint}(G,D_j) = \mathbb{E}_{x,y}[log(D_j(x,y))] + \mathbb{E}_{x,c}[log(1-D_j(x,G(x,c)))] \tag{1}$$

$$\mathcal{L}_{marg}(G,D_m) = \mathbb{E}_{x}[log(D_m(x))] + \mathbb{E}_{x,c}[log(1-D_m(G(x,c)))] \tag{2}$$

The learned GAN feature loss (3) is computed between the generated image and the corresponding target image. Here, $D_j^{(i)}$ denotes outputs of intermediate layers of the joint discriminator where T is the total number of intermediate layers. N_i denotes the number of elements in each layer and v_i are the assigned weights. These weights are monotonically decreasing, meaning that we put greater emphasis on on lower-level features.

$$\mathcal{L}_{feat}(G,D_j) = \mathbb{E}_{x,y,c}[\sum_{i=1}^{T} v_i \frac{1}{N_i} \left\| D_j^{(i)}(x,y) - D_j^{(i)}(x,G(x,c)) \right\|_1] \tag{3}$$

Conversely, a perceptual VGG loss (4) is applied between the generated image and the input, where the weights w_i are a monotonically increasing. By assigning greater weight to higher-level features, we hope to preserve image semantics such as cars and pedestrians. In this equation $F^{(i)}$ denotes intermediate layers of a pretrained VGG19 network where U is the total number of layers that are used for feature extraction. M_i denotes the number of elements in each layer.

$$\mathcal{L}_{vgg}(G) = \mathbb{E}_{x,c}[\sum_{i=1}^{U} w_i \frac{1}{M_i} \left\| F^{(i)}(x) - F^{(i)}(G(x,c)) \right\|_1] \tag{4}$$

Similar to StarGAN [2] we introduce a domain classification loss (5) and (6), penalizing false domain predictions of the fake and real images, respectively. Since we have binary encoded labels, we use Binary Cross Entropy Loss.

$$\mathcal{L}^f_{cls}(G, D_m) = -\mathbb{E}_{x,c}[c\, log(D_{m,cls}(G(x,c))) + (1-c)\, log(1 - D_{m,cls}(G(x,c)))] \tag{5}$$

$$\mathcal{L}^r_{cls}(D_m) = -\mathbb{E}_{x,c}[c\, log(D_{m,cls}(x)) + (1-c)\, log(1 - D_{m,cls}(x))] \tag{6}$$

The regularization in (7) for the attention mask $M(x,c)$ is chosen so that the generated mask never fully reaches zero, since that would make the mask redundant. Note, that this is a monotonically decreasing function. To smooth out the mask a total variation loss (8) is applied.

$$\mathcal{L}_{att}(G) = \mathbb{E}_{x,c}[log(\|M(x,c)\|_1) + \frac{1}{\|M(x,c)\|_1}] \tag{7}$$

$$\mathcal{L}_{TV}(G) = \mathbb{E}_{x,c}[\sum_{i,j=1}^{H,W} [(M_{i+1,j}(x,c) - M_{i,j}(x,c))^2 + (M_{i,j+1}(x,c) - M_{i,j}(x,c))^2]] \tag{8}$$

Full Objective. The full training objectives for G, D_j and D_m are given below.

$$\mathcal{L}_G = \mathcal{L}_{joint} + \lambda_{marg}\mathcal{L}_{marg} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{vgg}\mathcal{L}_{vgg} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{att}\mathcal{L}_{att} + +\lambda_{TV}\mathcal{L}_{TV} \tag{9}$$

$$\mathcal{L}_D = \mathcal{L}_{D_j} + \mathcal{L}_{D_m} = \mathcal{L}_{joint} + \lambda_{marg}\mathcal{L}_{marg} + \lambda_{cls}\mathcal{L}^r_{cls} \tag{10}$$

1.2 Training Specifications

All models were trained for 30 epochs on a single GPU for about 4 days. The learning rate is constant for 20 epochs and linearly decayed to 0 for the following 10 epochs. The batch size is 6, λ_{marg} = 100.0, λ_{feat} = 100.0, λ_{vgg} = 10.0, λ_{cls} = 1.0, λ_{att} = 1.0, λ_{TV} = 0.0001. Image size is (256x256). Other model parameters were set to default according to SPADE. With 4 upsampling layers, the generator has around 100 million parameters, whereas the discriminator has around 30 million.

When training the model we noticed that the GAN feature loss seemed to be very unstable. The authors of SPADE argue that this loss was crucial in generating good looking samples. Indeed, when disabling this loss the generated images get significantly worse.

2 More Image Translation Results

Figure 1 shows several failure cases where the input domain is WNO. The model is able to capture the different styles of each domain. However, it produces severe artifacts when translating images from night to day. Since the domain conditioning of our model consists of binary encoded labels, we can arbitrarily create any sort of combinations. Figure 2 shows results from these "mixed domains" where the class labels are (SNS - Summer, Night, Sunny), a paradoxical domain, and (WDR - Winter, Day, Rain). Interestingly, the model is able to generate samples from WDR despite never seeing samples from this domain during training. This suggests that our model may have an understanding of each individual attribute within the class label. Lastly, additional images from other ablation models are shown in Figure 3.

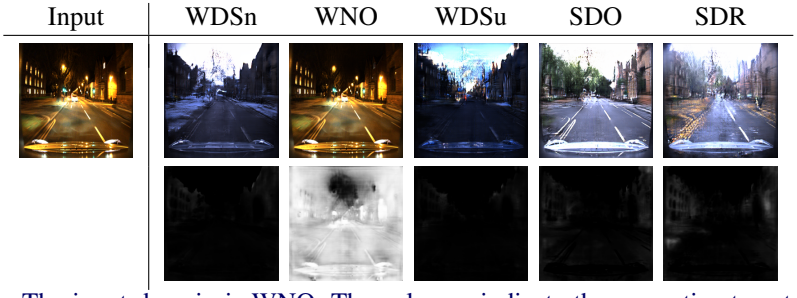

Figure 1: The input domain is WNO. The columns indicate the respective target domain, while the second row depicts the corresponding attention masks.

(a) Input (b) SNS (c) WDR (e) Input (f) SNS (g) WDR

Figure 2: Mixed Domain examples. Input domain is WDO for both cases.

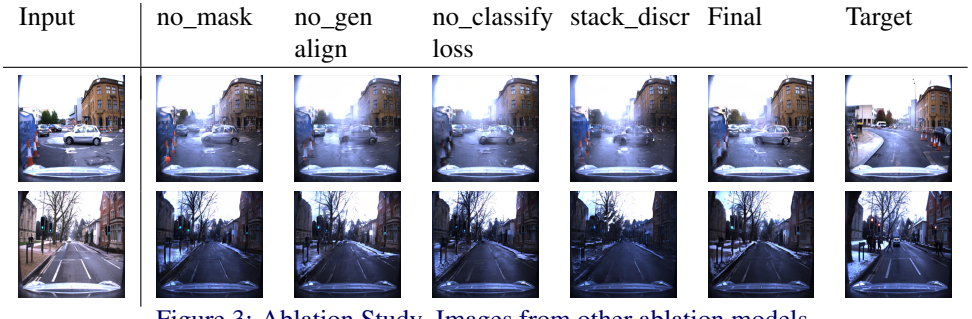

Figure 3: Ablation Study. Images from other ablation models.

3 More Image Retrieval Results

Figure 4 is obtained with the same experimental setup as Figure 7 in our paper. Figure 7 in the paper, however, reports the average retrieval accuracy over all query conditions, while Figure 4 shows the results for each query condition separately.

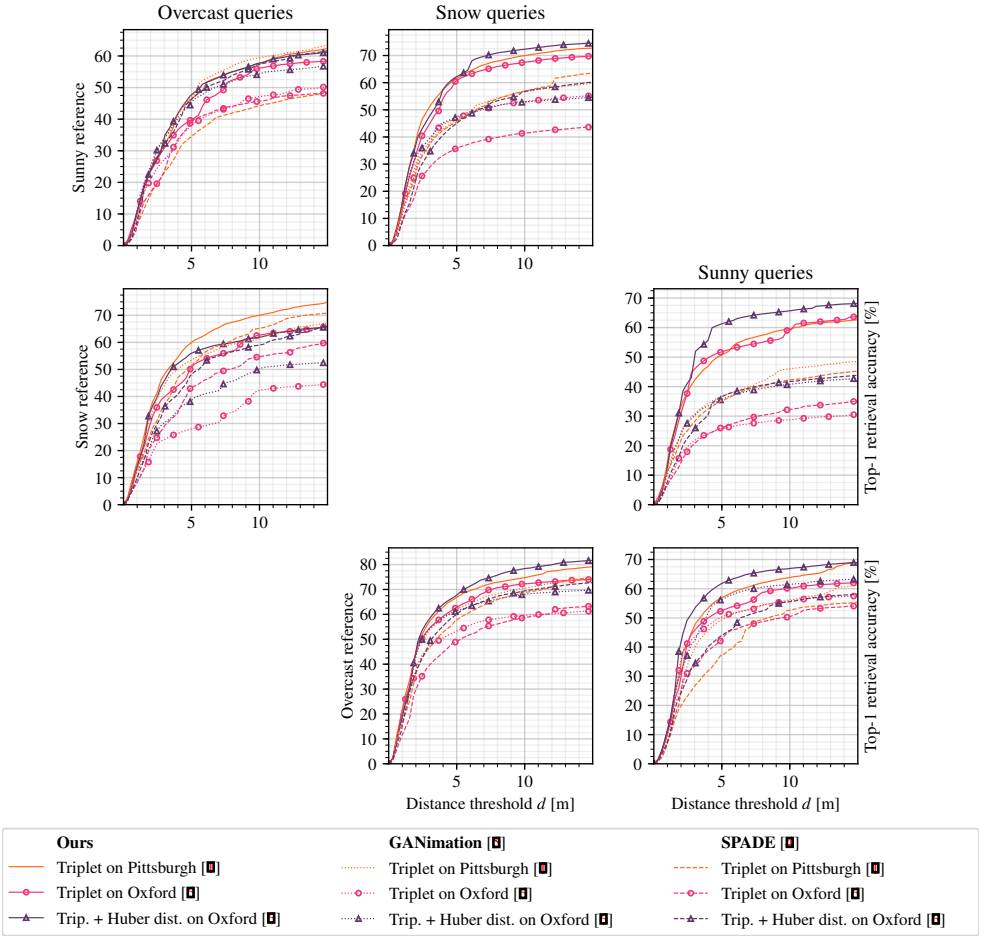

Figure 4: Top-1 retrieval accuracy on the Oxford RobotCar dataset for three different reference conditions (from top to bottom: sunny, snow and overcast) and three different query conditions (from left to right: overcast, snow and sunny) as a function of the distance threshold d for correct retrieval. 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 [6] and SPADE [4] by combining each with three different image retrieval networks from [0] and [3] that have been trained either on the Pittsburgh [7] dataset or the Oxford RobotCar [3] dataset.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.