Exemplar Guided Unsupervised Image-to-Image Translation

Liqian Ma[1] **Xu Jia**[2] **Stamatios Georgoulis**[1,3] **Tinne Tuytelaars**[2] **Luc Van Gool**[1,3]
[1]KU-Leuven/PSI, TRACE (Toyota Res in Europe) [2]KU-Leuven/PSI, IMEC [3]ETH Zurich
{liqian.ma, xu.jia, tinne.tuytelaars, luc.vangool}@esat.kuleuven.be
{georgous, vangool}@vision.ee.ethz.ch

Abstract

Image-to-image translation task has become a popular topic recently. Most works focus on either one-to-one mapping in an unsupervised way or many-to-many mapping in a supervised way. However, a more practical setting is many-to-many mapping in an unsupervised way, which is harder due to the lack of supervision and the complex inner and cross-domain variations. To alleviate these issues, we propose the Exemplar Guided UNsupervised Image-to-image Translation (EG-UNIT) network which conditions the image translation process on an image in the target domain. An image representation is assumed to comprise both content information which is shared across domains and style information specific to one domain. By applying exemplar-based adaptive instance normalization to the shared content representation, EG-UNIT manages to transfer the style information in the target domain to the source domain. Experimental results on various datasets show that EG-UNIT can indeed translate the source image to diverse instances in the target domain with semantic consistency.

1 Introduction

Image-to-image translation has received significant attention, as it enables numerous applications, such as image editing [34, 34], attribute/style transfer [3, 29] and domain adaptation [12, 26]. The goal in each case is to map an image from a source domain to a target domain. For example, semantic maps to real images, gray-scale to color images, low-resolution to high-resolution images, and so on. Recent works [15, 34, 38] have been very successful in translating images from one domain to another using paired data as supervision. However, for many tasks it is not easy or even possible to obtain such paired data, *e.g*. in cross-city street view translation or male-female face translation. In the unsupervised setting, there are no paired data to show how an image in the source domain should be translated to an image in the target domain. For this setting, Zhu *et al*. [37] proposed a cycle-consistency loss which assumes that a mapping followed by its reverse operation approximately yields an identity function, that is, $F(G(x_{\mathrm{A}})) \approx x_{\mathrm{A}}$. Liu *et al*. [26] further proposed a shared-latent space constraint which assumes that a pair of corresponding images $(x_{\mathrm{A}}, x_{\mathrm{B}})$ in two different domains, A and B, can be mapped to the same representation z in a shared latent space Z. However, all the aforementioned methods assume that there is a deterministic one-to-one mapping between the two domains, *i.e*. each image is translated to only a single output image. By doing so, they fail to capture the diversity of the image distribution within the target domain, such as different color and style of shoes in sketch-to-image translation and different seasons in synthetic-to-real street view translation.

Inspired by Bousmalis *et al*. [2], we propose the EG-UNIT framework to address this issue. The core idea is that an image representation is composed of i) a representation shared between the two domains, which models the content in the image, and ii) a representation which contains style information specific to that domain. However, for a domain with complex inner-variations, *e.g*. street views of day-and-night or different seasons, it is difficult to have a single static representation which

Preprint. Work in progress.

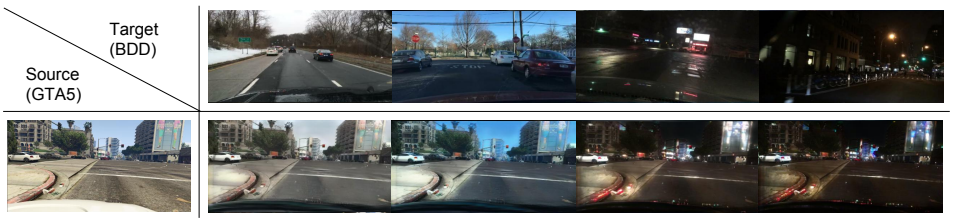

Figure 1: Exemplar guided image translation examples of GTA5 → BDD. Best viewed in color.

covers all variations in that domain. Moreover, it is unclear which style (time-of-day/season) to pick during the image translation process. Therefore, we propose to condition the image translation process on an arbitrary image in the target domain. Using different image exemplars as guidance, the proposed method is able to translate an input image into images of different styles within the target domain. Therefore, our framework enables many-to-many image translations and allows explicit control over the translation process to produce images of desired styles – see Fig. 1.

To instantiate this idea, we build our EG-UNIT framework on top of the weight sharing architecture proposed in UNIT [26]. However, instead of having a single latent space which is shared by both domains, we propose to decompose the latent space into two components. One component is shared by the two domains and focusses on the image content, while the other captures the style information associated with the exemplars in the target domain. To translate an image with an exemplar in the target domain as guidance, we apply adaptive instance normalization (AdaIN) [13] to each image's shared content representation with parameters computed from the target domain exemplar. The shared content representation contains semantic information, such as the objects' category, shape and spatial layout, while the adaptive instance normalization is able to transfer the style information, such as the color and texture, from a target domain exemplar to an image in the source domain. However, directly applying AdaIN to feature maps of content representation would mix up all objects and scenes in the image and the image translation would be prone to failure when an image contains diverse objects and scenes. To avoid this issue, previous style transfer works [8, 25, 28] take a semantic label map as additional input. However, semantic label maps are not easy to obtain for most tasks. To maintain semantic consistency during image translation, we propose to compute feature masks to approximately decouple different semantic categories in an unsupervised way with the help of perceptual losses and adversarial loss. Ideally, one feature mask corresponding to a certain category is applied to one feature map and the adaptive instance normalization for that channel is only required to capture and model the difference for that category, *e.g.* sky's style in two domains. See Fig. 2 for an overview of the EG-UNIT architecture.

Our contribution is two-fold. i) We propose a novel approach EG-UNIT for the image-to-image translation task, which enables many-to-many mappings and allows explicit style control over the translation process. ii) Evaluation on both designed controlled datasets and street view datasets show that our method is robust to mode collapse and can generate semantically consistent results conditioned on a given exemplar image.

2 Related work

Image-to-image translation. Image-to-image translation is used to learn a mapping from one image (*i.e.* source domain) to another (*i.e.* target domain). Recently, with the advent of generative models, such as [11, 20], there have been a lot of works on this topic. Isola *et al.* proposed pix2pix [16] to learn the mapping from input images to output images using a U-Net neural network in an adversarial way. Wang *et al.* extended the method to pix2pixHD [34], to turn semantic label maps into high-resolution photo-realistic images. Zhu *et al.* extended pix2pix to BicycleGAN [38], which can model multimodal distributions and produce both diverse and realistic results. All these methods, however, require paired training data as supervision which may be difficult or even impossible to collect in many scenarios such as synthetic-to-real street view translation or face-to-cartoon translation [32].

Recently, several unsupervised methods have been proposed to learn the mappings between two image collections without paired labels. However, this is an ill-posed problem because there are infinitely many mappings existing between two unpaired image domains. To address this ill-posed problem, different constraints have been added to the network to regularize the learning process [18,

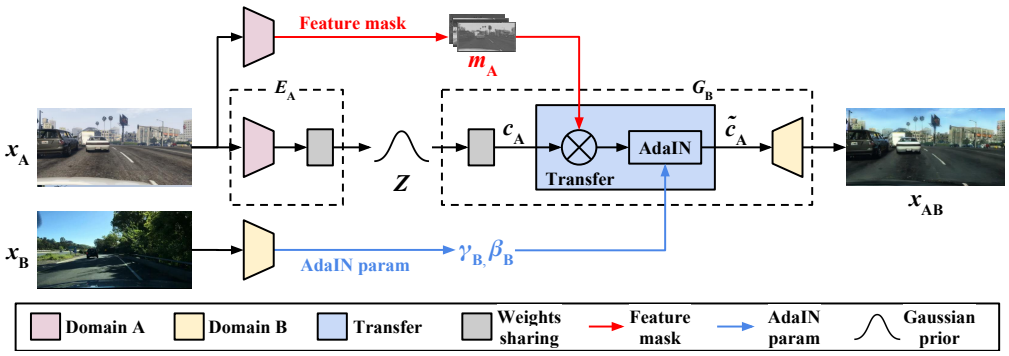

Figure 2: The x_A to x_{AB} translation procedure of our EG-UNIT framework. 1) Source domain image x_A is fed into an encoder E_A to compute a shared latent code z_A and is further decoded to a common high-level content representation c_A. 2) Meanwhile, x_A is also fed into a sub-network to compute feature masks m_A. 3) The target domain exemplar image x_B is fed to a sub-network to compute affine parameters γ_B and β_B for AdaIN . 4) The content representation c_A is transferred to the target domain using m_a, γ_B, β_B, and is further decoded to an image x_{AB} by target domain generator G_B.

26, 32, 36, 37]. One popular constraint is cycle-consistency, which enforces the network to learn deterministic mappings for various applications [18, 36, 37]. Going one step further, Liu *et al*. [26] proposed a shared-latent space constraint which encourages a pair of images from different domains to be mapped to the same representation in the latent space. Royer *et al*. [32] proposed to enforce a feature-level constraint with a latent embedding reconstruction loss. However, we argue that these constraints are not well suited for complex domains with large inner-domain variations, as also mentioned in [1, 14]. To address this problem, we propose to add a target domain exemplar as guidance in image translation. This allows many-to-many translation and can produce images of desired styles with explicit control. The style information is transferred from the exemplar image to the content image through AdaIN [13]. Concurrent to our work, Huang *et al*. [14], also proposed to use AdaIN to transfer style information from the target domain to the source domain. However, before applying AdaIN to the shared content representation, we compute feature masks to decouple different semantic categories. By applying feature masks to the feature maps in the shared content space, each channel can specialize and model the difference for a single category only, which helps handle a domain with complex scenes.

Style transfer. Style transfer aims at transferring the style information from the exemplar image to the content image, while preserving the content information. The seminal work by Gatys *et al*. [7] proposed to transfer style information by matching the feature correlations, *i.e*. Gram matrices, in the convolutional layers of a deep neural network (DNN) following an iterative optimization process. In order to improve the speed and flexibility, several feed-forward neural networks have been proposed [6, 10, 13, 17, 23, 24]. Huang *et al*. [13] proposed a simple but effective method, called adaptive instance normalization (AdaIN), which aligns the mean and variance of the content image features with those of the style image features. Li *et al*. [24] proposed the whitening and coloring transform (WCT) algorithm, which directly matches the features' covariance in the content image to those in the given style image. However, due to the lack of semantic consistency, these stylization methods usually generate non-photo-realistic images with the "spills over" problem [28]. To address this, semantic label maps are used as additional input to help style transfer between corresponding semantic regions [9, 25, 28]. In this paper, we propose to compute feature masks to approximately model such semantic information in an unsupervised way.

3 Method

Our goal is to learn a many-to-many mapping between two domains in an unsupervised way. For example, a synthetic street view image can be translated to either a day-time or night-time realistic scene. We assume that an image representation can be decomposed into two components, one modeling the shared content between domains and the other modeling style information specific to exemplars in the target domain. The style information within an exemplar image in the target domain is transferred to an image in the source domain through an adaptive instance normalization layer.

3.1 Framework

For simplicity, we present EG-UNIT in the A→B direction – see Fig. 2. It includes an encoder E_{A}, a generator G_{A}, and a discriminator D_{A}. For the B→A direction, the translation process is analogous.

Weight sharing for common content representation. The weight sharing strategy is proposed in UNIT [26] to implicitly model the cycle-consistency constraint. The latter assumes that any matched image pair in the two domains can be mapped to the same latent representation in a shared-latent space. However, the translation in UNIT is done by directly decoding the common latent representation to an image in the target domain, which restricts the method to a one-to-one mapping. In our work, we also adopt the weight sharing strategy, but only use it to compute a common latent representation for content, c_{A}. Style information is injected into it at the decoding stage, which allows for a many-to-many mapping as will be explained later. Each image domain (*i.e.* source and target) is modeled by a VAE-GAN [21], which contains an encoder E, a generator G, and a discriminator D. Common content representation can be computed in two steps. First, E_{A} encodes the source domain image x_{A} into a latent code z_{A} in the shared-latent space. Second, a random-perturbed version of the latent code z_{A} is decoded to a high-level feature c_{A} representing the domain-common content information. More specifically, the network inferences a mean vector $E_{\mu}(x_{\mathrm{A}})$, and the distribution of the latent code z_{A} is given by $q(z_{\mathrm{A}}|x_{\mathrm{A}}) \equiv N(z_{\mathrm{A}}|E_{\mu}(x_{\mathrm{A}}), I)$ where I is an identity matrix. For more details about the VAE-GAN and weight-sharing technique, we refer the reader to [26].

Exemplar-based Adaptive Instance Normalization. The shared content representation contains semantic information but not style information. In [13], it is shown that affine parameters in instance normalization have a big influence on the output image's style. Therefore, we propose to apply instance normalization to the shared content representation before the decoding stage. An image in the target domain is fed to another network to compute a set of feature maps f_{B}, which are expected to contain the style information of the target domain. Similar to [13], means and variances are calculated for each channel of f_{B} and used as the affine parameters in instance normalization, as shown in Eq. 1,

$$\gamma_{\mathrm{B}} = \mu(f_{\mathrm{B}}), \quad \beta_{\mathrm{B}} = \delta(f_{\mathrm{B}}), \tag{1}$$

$$\mathrm{AdaIN}(c_{\mathrm{A}}, \gamma_{\mathrm{B}}, \beta_{\mathrm{B}}) = \gamma_{\mathrm{B}} \frac{(c_{\mathrm{A}} - \mu(c_{\mathrm{A}}))}{\delta(c_{\mathrm{A}})} + \beta_{\mathrm{B}}, \tag{2}$$

where $\mu(\cdot)$ and $\delta(\cdot)$ respectively denote a function to compute the mean and variance across spatial dimensions. The shared content representation is first normalized by these affine parameters as shown in Eq. 2 and then decoded to a target-domain image with a target domain generator. Since different affine parameters normalize the feature statistics in different ways, by using different exemplar images in the target domain as input we can translate an image in the source domain to different sub-styles in the target domain. Therefore, EG-UNIT allows for multi-modal image-to-image translations and allows the user to have explicit style control over the translation process.

However, directly applying such adaptive instance normalization to the shared content representation does not give satisfying results. The reason is that one channel in the shared content representation is likely to contain information from multiple objects and scenes. The difference of these objects and scenes between the two domains is not always uniform. Applying a normalization over a feature map with complex semantics is prone to mix styles of different objects and scenes together, hence failing to give semantic-consistent translation. We propose to compute feature masks to make an approximate estimation about semantic categories. The feature masks m_{A} are computed by applying a nonlinear activation function and a threshold to feature maps f_{A}, as shown in Eq. 3,

$$m_{\mathrm{A}} = (1 - \eta) \cdot \sigma(f_{\mathrm{A}}) + \eta, \tag{3}$$

where η is a threshold and σ is the sigmoid function. Feature masks contain substantial semantic information, which can be used to keep the semantic consistency during the translation, *e.g.* translating the source sky into the target sky style without affecting the other scene elements. The new normalized representation $\tilde{c}_{\mathrm{A}}$ is computed using Eq. 4,

$$\tilde{c}_{\mathrm{A}} = \mathrm{AdaIN}(m_{\mathrm{A}} \circ c_{\mathrm{A}}, \gamma_{\mathrm{B}}, \beta_{\mathrm{B}}), \tag{4}$$

where $\circ$ denotes the Hadamard product.

During training, there are four types of information flow, as shown in Fig. 3. For the reconstruction flow $x_{\mathrm{A}} \rightarrow x_{\mathrm{AA}}$, content representation c_{A}, feature masks m_{A}, and AdaIN parameters $\gamma_{\mathrm{A}}, \beta_{\mathrm{A}}$ are

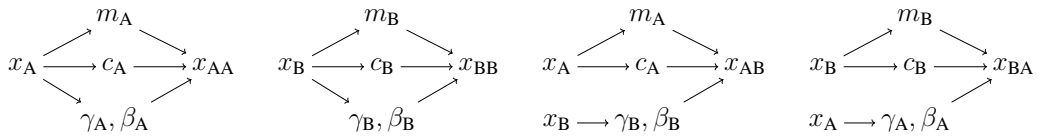

Figure 3: Information flow diagrams of auto-encoding procedures $x_{\mathrm{A}} \rightarrow x_{\mathrm{AA}}$ and $x_{\mathrm{B}} \rightarrow x_{\mathrm{BB}}$, and translation procedures $x_{\mathrm{A}} \rightarrow x_{\mathrm{AB}}$ and $x_{\mathrm{B}} \rightarrow x_{\mathrm{BA}}$.

all computed from x_{A} (and vice versa for $x_{\mathrm{B}} \rightarrow x_{\mathrm{BB}}$). For the translation flow $x_{\mathrm{A}} \rightarrow x_{\mathrm{AB}}$, content representation c_{A} and feature masks are computed from x_{A}, while AdaIN's affine parameters γ_{B} and β_{B} are computed from the target domain exemplar x_{B} (and vice versa for $x_{\mathrm{B}} \rightarrow x_{\mathrm{BA}}$).

3.2 Network architecture

The overall framework can be divided into several subnetworks[1]. 1) Two Encoders, E_{A} and E_{B}. Each one consists of several strided convolutional layers and several residual blocks to compute shared content representation; 2) A feature mask network and an AdaIN network, GB_{A} and GB_{B}, have the similar architecture as the Encoder except for the weight-sharing layers. 3) Two Generators, G_{A} and G_{B}, are almost symmetric to the Encoders except that the up-sampling is done by transposed convolutional layers. 4) Two Discriminators, D_{A} and D_{B}, are fully-convolutional networks containing a stack of convolutional layers. 5) VGG sub-network [33], VGG, contains the first few layers (up to relu5_1) of a pre-trained VGG-19 [33], which is used to calculate perceptual losses.

3.3 Learning

The learning procedure of EG-UNIT contains VAE, GAN and perceptual losses. To make the training more stable, we first pre-train the feature mask network and AdaIN network for each domain separately within a VAE-GAN architecture and use the encoder part as fixed feature extractors, *i.e.* GB_{A} and GB_{B}, for the following training.

The overall loss, shown in Eq. 5, consists of the UNIT loss $\mathcal{L}_{\mathrm{UNIT}}$ [26] and the perceptual loss $\mathcal{L}_{\mathrm{P}}$:

$$\min_{E_{\mathrm{A}},E_{\mathrm{B}},G_{\mathrm{A}},G_{\mathrm{B}}} \max_{D_{\mathrm{A}},D_{\mathrm{B}}} \mathcal{L}_{\mathrm{UNIT}}(E_{\mathrm{A}}, G_{\mathrm{A}}, D_{\mathrm{A}}, E_{\mathrm{B}}, G_{\mathrm{B}}, D_{\mathrm{B}}) + \mathcal{L}_{\mathrm{P}}(E_{\mathrm{A}}, G_{\mathrm{A}}, E_{\mathrm{B}}, G_{\mathrm{B}}) \tag{5}$$

The UNIT loss consists of VAE losses, GAN losses, and cycle-consistency losses. The VAE, GAN and cycle-consistency losses are identical to the ones used in [26].

$$\begin{aligned}\mathcal{L}_{\mathrm{UNIT}}(E_{\mathrm{A}}, G_{\mathrm{A}}, D_{\mathrm{A}}, E_{\mathrm{B}}, G_{\mathrm{B}}, D_{\mathrm{B}}) =& \mathcal{L}_{\mathrm{VAE_A}}(E_{\mathrm{A}}, G_{\mathrm{A}}) + \mathcal{L}_{\mathrm{GAN_A}}(E_{\mathrm{A}}, G_{\mathrm{A}}, D_{\mathrm{A}}) + \mathcal{L}_{\mathrm{CC_A}}(E_{\mathrm{A}}, G_{\mathrm{A}}, E_{\mathrm{B}}, G_{\mathrm{B}}) \\ &+ \mathcal{L}_{\mathrm{VAE_B}}(E_{\mathrm{B}}, G_{\mathrm{B}}) + \mathcal{L}_{\mathrm{GAN_B}}(E_{\mathrm{B}}, G_{\mathrm{B}}, D_{\mathrm{B}}) + \mathcal{L}_{\mathrm{CC_B}}(E_{\mathrm{A}}, G_{\mathrm{A}}, E_{\mathrm{B}}, G_{\mathrm{B}})\end{aligned} \tag{6}$$

Similar to [7, 17], our perceptual loss consists of the content loss captured by VGG feature map ϕ containing localized image spatial information, and the style loss captured by the Gram matrix containing non-localized image style information. The total perceptual loss is then

$$\begin{aligned}\mathcal{L}_{\mathrm{P}}(E_{\mathrm{A}}, G_{\mathrm{A}}, E_{\mathrm{B}}, G_{\mathrm{B}}) =& \lambda_c \mathcal{L}_{c_{\mathrm{A}}}(E_{\mathrm{A}}, G_{\mathrm{A}}) + \lambda_s \mathcal{L}_{s_{\mathrm{A}}}(E_{\mathrm{A}}, G_{\mathrm{A}}) \\ &+ \lambda_c \mathcal{L}_{c_{\mathrm{B}}}(E_{\mathrm{B}}, G_{\mathrm{B}}) + \lambda_s \mathcal{L}_{s_{\mathrm{B}}}(E_{\mathrm{B}}, G_{\mathrm{B}}),\end{aligned} \tag{7}$$

where λ_c and λ_s are the weights for content and style losses, which depend on the dataset domain variations and tasks. The content loss $\mathcal{L}_{c_{\mathrm{A}}}(E_{\mathrm{A}}, G_{\mathrm{A}})$ and style loss $\mathcal{L}_{s_{\mathrm{A}}}(E_{\mathrm{A}}, G_{\mathrm{A}})$ are defined as follows. $\mathcal{L}_{c_{\mathrm{B}}}(E_{\mathrm{B}}, G_{\mathrm{B}})$ and $\mathcal{L}_{s_{\mathrm{B}}}(E_{\mathrm{B}}, G_{\mathrm{B}})$ are defined likewise.

$$\mathcal{L}_{c_{\mathrm{A}}}(E_{\mathrm{A}}, G_{\mathrm{A}}) = \sum_{l=1}^{L} \frac{l}{LC_l H_l W_l} \|\phi_l(x_{\mathrm{AB}}) - \phi_l(x_{\mathrm{A}})\|_1, \tag{8}$$

$$\mathcal{L}_{s_{\mathrm{A}}}(E_{\mathrm{A}}, G_{\mathrm{A}}) = \sum_{l=1}^{L} \frac{1}{C_l H_l W_l} \|GM_l(x_{\mathrm{AB}}) - GM_l(x_{\mathrm{B}})\|_1, \tag{9}$$

[1]More details about the parameters of the network architecture are given in the supplementary material.

Table 1: SSIM evaluation for single-digit translation. Higher is better.

	CycleGAN	UNIT	EG-UNIT w/o feature mask	EG-UNIT w/o AdaIN	EG-UNIT w/o $\mathcal{L}_{\text{P}}$	EG-UNIT (ours)
A $\rightarrow$ B	0.214	0.178	0.395	0.208	0.286	0.478
B $\rightarrow$ A	0.089	0.074	0.133	0.080	0.093	0.232

where L is the total number of convolutional layers and l indicates the l-th convolutional layer of the VGG network. We use ϕ_l to denote the feature map with shape $C_l \times H_l \times W_l$ extracted by the l-th convolutional layer. $GM_l(x_{\text{AB}})$ and $GM_l(x_{\text{B}})$ are the Gram matrices of the l-th convolutional layers. For the content losses $\mathcal{L}_{c_{\text{A}}}$ and $\mathcal{L}_{c_{\text{B}}}$, a linear weighting scheme is adopted to help the network focus more on the high-level semantic information rather than low-level color information. In both content and style losses we use the L1 distance, which performs better than the L2 in our experiments.

4 Experiments

We evaluate EG-UNIT's translation ability, *i.e*. how well it generates domain-realistic looking images, both qualitatively and quantitatively on three tasks with progressively increasing visual complexity: 1) single-digit translation; 2) multi-digit translation; 3) street view translation. We first perform an ablation study on various components of the proposed approach on the basic single-digit translation task. Then, we present results on more challenging translation tasks, and evaluate our method quantitatively on the semantic segmentation task. Finally, we apply EG-UNIT to the face gender translation task. Upon acceptance, the code and datasets will be made publicly available.

Single-digit translation. We set up a controlled experiment on the MNIST-Single dataset which is created based on the handwritten digits dataset MNIST [22]. The MNIST-Single dataset consists of two different domains as shown in Fig. 4. For domain A of both training/test sets, the foreground and background are randomly set to *black* or *white* but different from each other. For domain B of training set, the foreground and background for digits from 0 to 4 are randomly assigned a color from $\{$*red, green, blue*$\}$, and the foreground and background for digits from 5 to 9 are fixed to *red* and *green*, respectively. For domain B of testing set, the foreground and background of all digits are randomly assigned a color from $\{$*red, green, blue*$\}$ Such data imbalance is designed on purpose to test the translation diversity and generalization ability. As for diversity, we want to check whether a method would suffer from the mode collapse issue and translate the images to the dominant mode, *i.e*. *(red, green)*. As for generalization, we want to check whether the model can be applied to new styles in the target domain that never appear in the training set, *e.g*. translate number 6 from *black* foreground and *white* background to *blue* foreground and *red* background.

We first analyze the importance of three main components of our framework, *i.e*. feature masks, AdaIN, and perceptual loss, on the MNIST-Single dataset. As shown in Fig. 4, our EG-UNIT can successfully transfer the source image into the style of the exemplar image. Ablating the feature mask from our framework leads to incorrect foreground and background shape, indicating that feature masks can indeed provide semantic information to transfer the corresponding local regions. Without AdaIN, the network suffers from the mode collapse issue in A$\rightarrow$B translation, *i.e*. , all samples are transferred to the dominant mode with red foreground and $green$ background. The latter indicates that the exemplar's style information can help the network to learn many-to-many mappings and avoid the mode collapse issue. Without perceptual losses $\mathcal{L}_{\text{P}}$, colors of foreground and background are incorrect, which shows that perceptual losses can encourage the network to learn semantic knowledge, *i.e*. foreground and background, in an unsupervised way. As for other state-of-the-art methods, both CycleGAN [37] and UNIT [26] can only do deterministic image translation and suffer from mode collapse issue, such as *white* $\leftrightarrow$ *green* and *black* $\leftrightarrow$ *red* for CycleGAN as shown in Fig. 4. These qualitative observations are in accordance with the quantitative results as shown in Tab. 1, which indicates that our full EG-UNIT obtains significantly higher SSIM scores than other alternatives.

To verify EG-UNIT's ability to match the distribution of real target domain data and translated results, we visualize them using t-SNE embeddings [30] in Fig. 5. In general, our method can match the target domain distribution well, while others either collapse to few modes or mismatch the distributions.

Multi-digit translation. The MNIST-Multiple dataset is another controlled experiment designed to mimic the complexity in real-world scenarios. It can be utilized to test whether the network understands the semantics (*i.e*. , digits) in an image and translates each digit accordingly. Each image

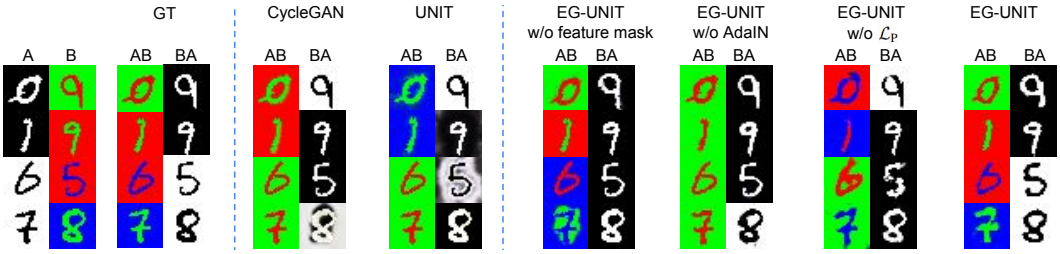

Figure 4: Single-digit translation testing results. The left-most four columns are samples from domain x_A and x_B, and reference translated ground truth x_{AB} and x_{BA}. Best viewed in color.

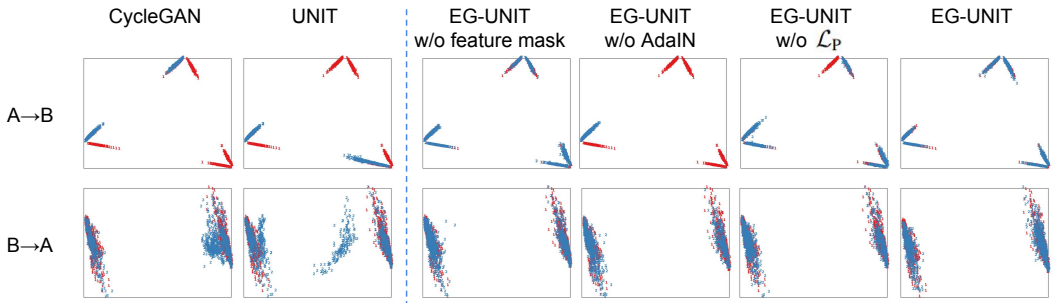

Figure 5: Single-digit translation t-SNE embedding visualization. Red: real samples. Blue: generated samples. Best viewed in color.

in MNIST-Multiple contains all ten digits, which are randomly placed in 4×4 grids. Two domains are designed: the foreground and background are randomly set to *black* or *white* but different from each other; for domain B, the background is randomly assigned to either *black* or *white* and each foreground digit is assigned to a certain color, but with a little saturation and lightness perturbation.

As seen in Fig. 6, both CycleGAN and UNIT can not translate the foreground and background color accordingly, and the colors in CycleGAN looks more "fake". This is due to the fact that CycleGAN and UNIT only learn a one-to-one mapping. These observations are consistent with the SSIM score in Tab. 2, where both CycleGAN and UNIT have low SSIM scores.

Table 2: SSIM score of multi-digit translation. Higher is better.

	CycleGAN	UNIT	EG-UNIT (ours)
A → B	0.145	0.239	0.503
B → A	0.130	0.234	0.495

Street view translation. We carry out a synthetic ↔ real experiment for street view translation on three datasets. 1) GTA5 [31], a synthetic dataset with 24,966 images from the game Grand Theft Auto V. 2) Cityscapes [5], a real-world dataset that has 5,000 images (2,975 train, 500 val, and 1,525 test) of ego-centric urban driving scenarios. 3) Berkeley Deep Drive (BDD) [35], a real-world dataset with 3,335 labeled training images and 19,110 unlabeled ones covering diverse driving scenes and lighting conditions. As we do not need any labels for translation, for training we mix BDD's labeled and unlabeled data. The street view datasets are more complex than the digit ones (different illumination/weather conditions, complex environments). The semantic annotations of all three datasets are compatible and contain 19 categories.

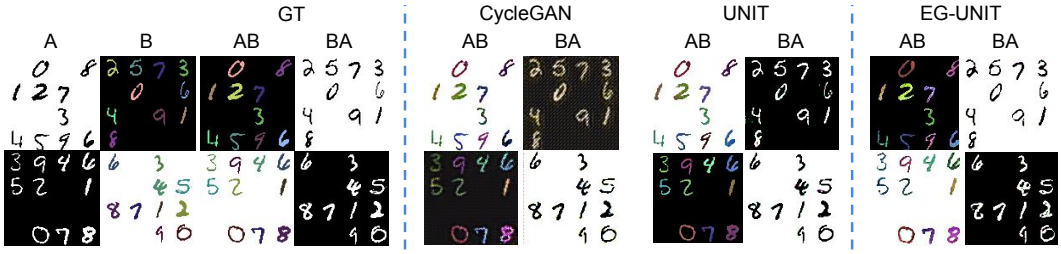

Figure 6: Multi-digit translation testing results. Best viewed in color.

Figure 7: Street view translation testing results of EG-UNIT. Best viewed in color.

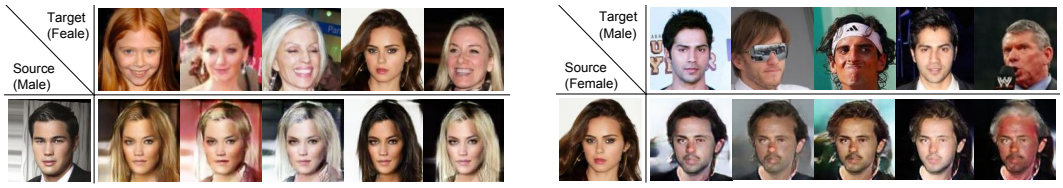

Figure 8: Face gender translation testing results of EG-UNIT. Best viewed in color.

Two sets of synthetic $\leftrightarrow$ real translation experiments are conducted: GTA5 $\leftrightarrow$ Cityscapes, and GTA5 $\leftrightarrow$ BDD. As shown in Fig. 7, our method can successfully translate the images from the source domain to the target domain according to the exemplar-specific style, even for a domain with large variations, *e.g.* day and night. For GTA5 $\leftrightarrow$ Cityscapes translation, the GTA5's road texture and scene style can be successfully translated. For the challenging GTA5 $\leftrightarrow$ BDD translation, the sky color and illumination condition can also be translated according to the exemplar image, even under large day/night variations. However, without feature mask, the network can only translate the image to day style even though a night exemplar image is given. Similar to FCN-score used in [15], we also use the performance of semantic segmentation task to quantitatively evaluate image translation quality as shown in Tab. 3. We first translate images in GTA5 dataset to an arbitrary image in Cityscapes or BDD dataset. We only generate images of size 256×512 due to the limitation on GPU memory. Then, we train a single-scale Deeplab model [4] on the translated images and test it on the validation set of Cityscapes and test set of BDD, respectively. The mIoU scores in Tab. 3 show that training with our translated synthetic images can improve the segmentation results, which indicates that our method can indeed reasonably translate the source GTA5 image to the target domain style.

Table 3: Semantic segmentation evaluation on 256×512 resolution.

	GTA → Cityscapes		GTA → BDD	
Method	mIoU	mIoU Gap	mIoU	mIoU Gap
Source	0.305	-0.225	0.329	-0.119
UNIT [26]	0.321	-0.209	0.297	-0.151
EG-UNIT (ours)	0.353	-0.177	0.343	-0.105
Oracle	0.530	0	0.448	0

Face gender translation. The Large-scale CelebFaces Attributes (CelebA) dataset [27] is a large-scale face attributes dataset with more than 200K celebrity images. We divide the aligned face images into male and female domains, containing 84,434 and 118,165 images respectively. We perform face gender translation on this dataset to show how the proposed method can be generalized to tasks with attributes as styles. From Fig. 8, we observe that EG-UNIT can translate the face gender successfully, and transfer the style of hair, skin and background according to the given exemplar image.

5 Conclusion

We introduce the EG-UNIT framework to learn a many-to-many mapping across domains in an unsupervised way. The image representation is assumed to be decomposed into a shared content representation and a style specific representation. EG-UNIT can transfer the style information of the target domain to the source domain with semantic consistency, by applying exemplar-based AdaIN to a shared content representation. Both quantitative and qualitative results demonstrate the effectiveness of our method.

Acknowledgments

We gratefully acknowledge the support of Toyota Motors Europe, FWO Structure from Semantics project, and KU Leuven GOA project CAMETRON.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, 2016.

[3] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *CVPR*, 2018.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017.

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016.

[8] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, 2017.

[9] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *CVPR*, 2017.

[10] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *BMVC*, 2017.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[12] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[13] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[14] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *arXiv preprint arXiv:1804.04732*, 2018.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, 2017.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 1412.6980, 2014.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *IJCAI*, 2017.

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NIPS*, 2017.

[25] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. *arXiv preprint arXiv:1802.06474*, 2018.

[26] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[28] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, 2017.

[29] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks (with supplementary materials). In *CVPR*, 2018.

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[31] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016.

[32] Amélie Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Moressi, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. *arXiv preprint arXiv:1711.05139*, 2017.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[34] 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. *arXiv preprint arXiv:1711.11585*, 2017.

[35] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *CVPR*, 2017.

[36] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017.

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[38] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017.

A Supplementary material

This supplementary material includes additional implementation details regarding the network architecture and training (§A.1), as well as comparisons with MUNIT [14] (§A.2).

A.1 Implementation details

The network architecture and training parameters are listed in Tab. 4. We set the number of down-sampling and up-sampling convolutional layer $n_1 = 1$ in single-digit translation and $n_1 = 3$ in other translation experiments. Following UNIT [26], the number of residual blocks in $Encoder$ and $Generator$ is set to $n_2 = 4$ with one sharing layer, and the number of convolutional layer discriminator is set to $n_3 = 5$. The threshold for computing feature mask is set to $\eta = 0.5$.

We use the Adam [19] optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is polynomially decayed with a power of 0.9, as mentioned in [4]. In order to keep training stable, we update encoder and generator 5 times, and discriminator 1 time in each iteration. The loss weights in $\mathcal{L}_{\text{UNIT}}$ are following [26], and λ_c, λ_s are chosen according to the dataset variations and tasks. For data augmentation, we do left-right flip and random corp. In addition, we set a low λ_c for face gender translation, since we need to change the shape and add/remove hair in this translation task.

Table 4: Network architecture and training parameters.

Translation	n_1	n_2	n_3	Minibatch	Learning rate	λ_s	λ_c	Iteration
Single-digit	1	4	5	8	1e-5	1e3	1e1	∼60k
Multi-digit	3	4	5	8	1e-5	1e4	1e2	∼60k
GTA5↔Cityscapes	3	4	5	3	1e-4	1e3	5e2	∼22k
GTA5↔BDD	3	4	5	3	1e-4	1e4	1e2	∼22k
Face gender	3	4	5	8	1e-4	5e3	1e1	∼30k

A.2 Comparisons with MUNIT

One concurrent work MUNIT [14] can also translate the source image with the exemplar's style. Here, we compare our EG-UNIT with MUNIT both qualitatively and quantitatively.

Single-digit translation. As shown in Fig.9, MUNIT can successfully transfer the style of the exemplar image, but the foreground and background are mixed and the digit's shape is not kept well in some cases. Our EG-UNIT can both transfer the style and keep the digit shape at the same time. This observation is verified by the SSIM score – see Tab. 5 – where EG-UNIT obtains higher SSIM scores, and the distribution visualization – see Fig. 10 – where EG-UNIT better matches the distribution between real and translated images.

Multi-digit translation. As shown in Fig. 9, MUNIT can successfully transfer the background color of the exemplar image, but fails in transferring the foreground digit color in A→B translation. This can also be observed from the distribution visualization in Fig. 11 where MUNIT mismatches the distribution in A→B translation.

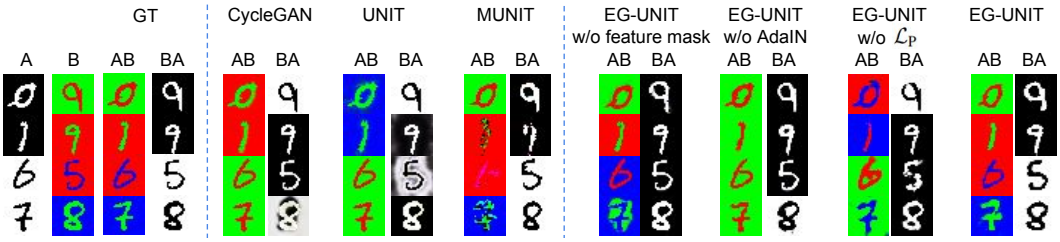

Figure 9: Single-digit translation testing results.

Table 5: SSIM evaluation for single-digit translation. Higher is better.

	CycleGAN	UNIT	MUNIT	EG-UNIT w/o feature mask	EG-UNIT w/o AdaIN	EG-UNIT w/o $\mathcal{L}_P$	EG-UNIT
A → B	0.214	0.178	0.463	0.395	0.208	0.286	0.478
B → A	0.089	0.074	0.227	0.133	0.080	0.093	0.232

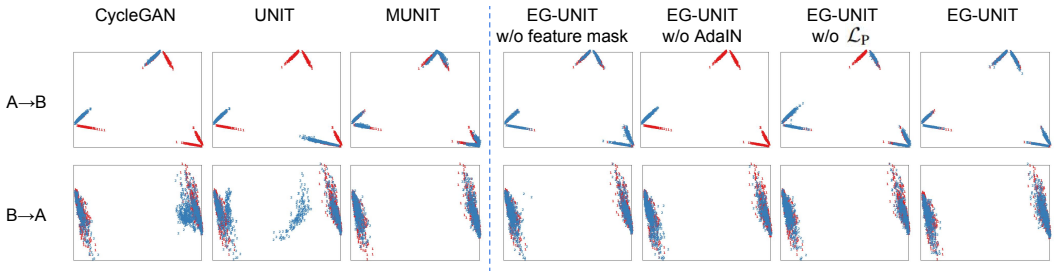

Figure 10: Single-digit translation t-SNE embedding visualization. Red: real samples. Blue: generated samples. Best viewed in color.

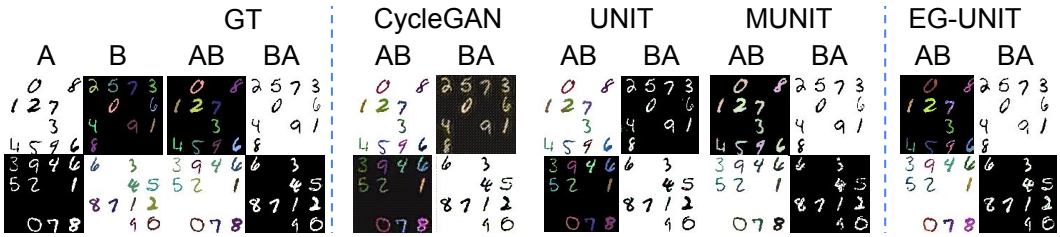

Figure 11: Multi-digit translation testing results.

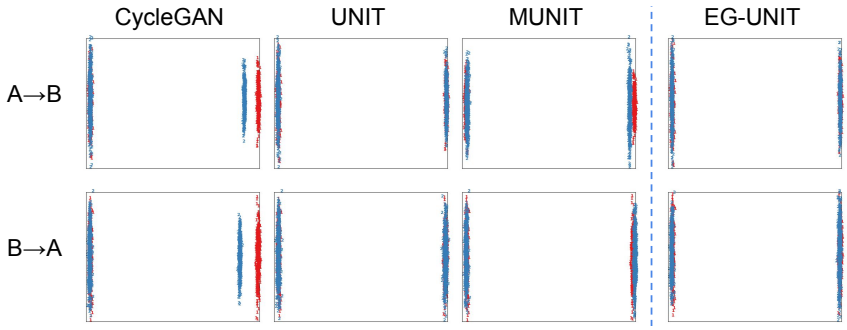

Figure 12: Multi-digit translation t-SNE embedding visualization. Red: real samples. Blue: generated samples. Best viewed in color.

Street view translation. As shown in Fig. 13, MUNIT tends to only succeed in day↔day translation but fails in day↔night translation. In contrast, our EG-UNIT can successfully translate the complex street view translation, even with large inner-domain variations, *e.g.* day/night. This naturally leads to higher mIoU scores on the semantic segmentation task, depicted in Tab. 6

Table 6: Semantic segmentation evaluation on 256×512 resolution.

	GTA → Cityscapes		GTA → BDD	
Method	mIoU	mIoU Gap	mIoU	mIoU Gap
Source	0.305	-0.225	0.329	-0.119
UNIT [26]	0.321	-0.209	0.297	-0.151
MUNIT [14]	0.345	-0.185	0.331	-0.117
EG-UNIT (ours)	0.353	-0.177	0.343	-0.105
Oracle	0.530	0	0.448	0

Face gender translation. As shown in Fig. 14, MUNIT can successfully translate the gender according to the given exemplar image, but fails to transfer the hair and skin style in some cases. In contrast, our EG-UNIT can successfully translate the hair, skin, background according to the given exemplar image.

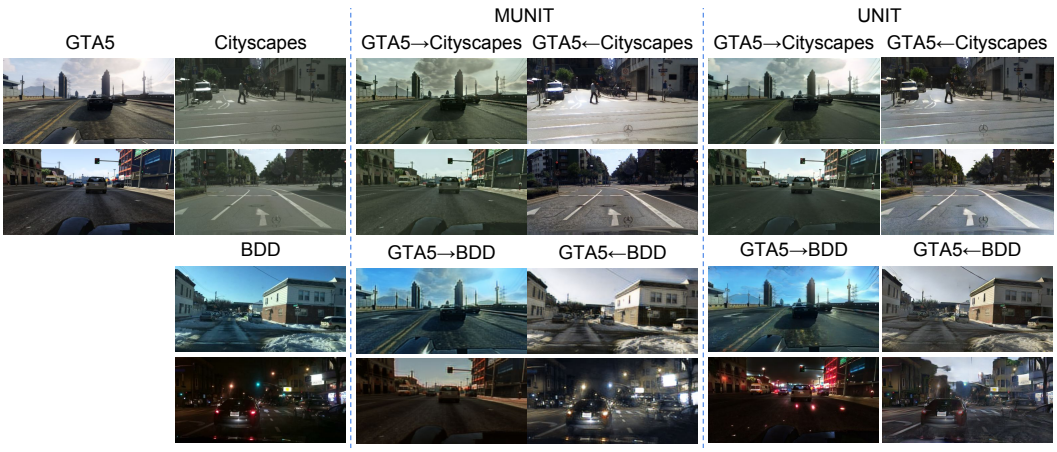

Figure 13: Street view translation testing results.

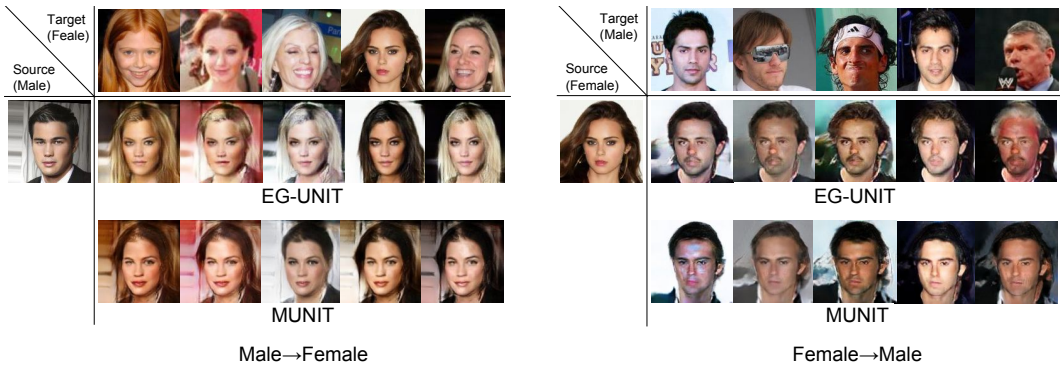

Figure 14: Face gender testing results.