Stochastic Actor-Executor-Critic for Image-to-Image Translation

Ziwei Luo[1*], **Jing Hu**[1*], **Xin Wang**[2†], **Siwei Lyu**[3], **Bin Kong**[2], **Youbing Yin**[2], **Qi Song**[2] and **Xi Wu**[1†]

[1]Chengdu University of Information Technology, China
[2]Keya Medical, Seattle, USA
[3]University at Buffalo, SUNY, USA
xi.wu@cuit.edu.cn, xinw@keyamedna.com

Abstract

Training a model-free deep reinforcement learning model to solve image-to-image translation is difficult since it involves high-dimensional continuous state and action spaces. In this paper, we draw inspiration from the recent success of the maximum entropy reinforcement learning framework designed for challenging continuous control problems to develop stochastic policies over high dimensional continuous spaces including image representation, generation, and control simultaneously. Central to this method is the *Stochastic Actor-Executor-Critic* (SAEC) which is an off-policy actor-critic model with an additional executor to generate realistic images. Specifically, the actor focuses on the high-level representation and control policy by a stochastic latent action, as well as explicitly directs the executor to generate low-level actions to manipulate the state. Experiments on several image-to-image translation tasks have demonstrated the effectiveness and robustness of the proposed SAEC when facing high-dimensional continuous space problems.

1 Introduction

Many computer vision problems, such as face inpainting, semantic segmentation, and realistic photo generated from sketch, can be defined as learning image-to-image translation (I2IT). Currently the most effective I2IT solution is based on a one-step framework that generates images in a single run of a deep learning (DL) model, such as VAE, U-Net, and conditional GANs. Directly learning I2IT with these DL models is challenging, due to the abundance of local minimums and poor generalization caused by overfitting. Although these problems could be potentially alleviated by using multi-scale models or a multi-stage pipelines, we are still left with models that have intrinsically high complexities, for which the optimal parameters (e.g. stage number and scale factor) have to be determined in a subjective and *ad hoc* manner.

To address these limitations of DL-based methods, we explore solving I2IT problems by leveraging the recent advances in deep reinforcement learning (DRL). The key idea is to decompose the monolithic learning process into small steps by a lighter-weight CNN, with the aim of progressively improving the quality of the model. Although recent works have successfully applied DRL to solve several visual tasks [Caicedo and Lazebnik, 2015], the action space in their tasks is usually discrete and can not be used for I2IT that requires continuous action spaces.

Currently, a promising direction for learning continuous actions is the maximum entropy reinforcement learning (MERL), which improves both exploration and robustness by maximizing a standard RL objective with an entropy term [Haarnoja *et al.*, 2018]. Soft actor-critic (SAC) [Haarnoja *et al.*, 2018] is an instance of MERL and has been applied to solve continuous action tasks. However, the main issue hindering the practical applicability of SAC on I2IT is its inability to handle high-dimensional states and actions effectively. Although recent work [Yarats *et al.*, 2019] addresses this problem by combining SAC with a regularized autoencoder (RAE), RAE only provides an auxiliary loss for an end-to-end RL training and is incapable for I2IT tasks.

Besides, a glaring issue is that the reward signal is so sparse that leads to an unstable training, and higher dimensional states such as pixels worsen this problem [Yarats *et al.*, 2019]. Thus, training an image-based RL model requires much more exploration and exploitation. One solution to stablize training is to extract a lower dimensional visual representation with a separately pre-trained DNN model, and learn the value function and corresponding policy in the latent spaces [Nair *et al.*, 2018]. However, this approach can not be trained from scratch, which could lead to inconsistent state representations with an optimal policy. Previous works like [Lee *et al.*, 2020] have shown that it is beneficial to learn the image representations and continuous controls simultaneously.

In this paper, we propose a new DRL architecture, *stochastic actor-executor critic* (SAEC), to handle I2IT with very high dimensional continuous state and action spaces. SAEC is formed by three core deep neural networks: the actor, the critic, and the executor (see Fig. 1). High level actions are generated by the actor in a low dimensional latent space, and forwarded to the executor for image translation. The critic evaluates the latent actions to guide the policy update. These

[*]Equal contribution.
[†]Corresponding authors.

three neural networks form two different branches in SAEC. More specifically, the actor and the critic form the RL branch as an actor-critic model, while The actor and executor form the DL branch similar to autoencoder, with an important difference that skip connections are added between the actor and executor. The RL branch is trained based on the MERL framework, while the DL branch is trained by minimizing a reconstruction objective.

Our contributions can be summarized as follows:

- A new DRL framework *stochastic actor-executor critic* (SAEC) is proposed to handle very high dimensional state and action spaces in I2IT task.
- Even using a sparse reward signal, our SAEC model can be stably trained from scratch on I2IT problems by combining DL-based supervision.
- Our SAEC framework is flexible to incorporate many advanced DL methods for various complex I2IT applications. And this framework enables the agent to simultaneously learn feature representation, control, and image generation in high-dimensional continuous spaces.

2 Background

2.1 Image-to-Image Translation

Image-to-image translation (I2IT) is often addressed by learning a generative process G that maps state $\mathbf{x}$ to target $\mathbf{y}$, $G : \mathbf{x} \rightarrow \mathbf{y}$. The state should be consistent with the target and both of them are images, such as generating realistic photos from semantic segmentation labels [Isola *et al.*, 2017], or synthesizing completed visually targets from images with missing regions [Pathak *et al.*, 2016]. Autoencoder is leveraged in most research work to learn this process by minimizing the reconstruction error $\mathcal{L}_{rec}$ between the predicted image $\tilde{\mathbf{y}}$ and the target $\mathbf{y}$ (see Figure 1). In addition, the generative adversarial network (GAN) is also vigorously studied in I2IT to synthesis realistic images [Isola *et al.*, 2017]. Subsequent works enhance I2IT performance by using a coarse-to-fine deep learning framework [Yu *et al.*, 2018] that recursively uses the previous stage's output as the input to the next stage (see the bottom of the Fig. 1(a)). In this way, I2IT task is transformed into a multi-stage, coarse-to-fine solution. Although iteration can be infinitely applied in theory, it is limited by the increasing model size and training instability.

2.2 Reinforcement Learning

Reinforcement learning (RL) is an infinite-horizon Markov decision process (MDP), defined by the tuple $(\mathcal{X}, \mathcal{A}, P, r, \gamma)$. $\mathcal{X}$ is a set of states, $\mathcal{A}$ is action space, and P represents the state transition probability given $x \in \mathcal{X}$ and $\mathbf{a}$, r is the reward emitted from each transition, and $\gamma \in [0, 1]$ is the reward discount factor. The standard objective of RL is to maximize the expected sum of rewards. Maximum entropy RL (MERL) further adds an entropy term: $\sum_{t=0}^{T} \mathbb{E}_{(\mathbf{x}_t, \mathbf{a}_t) \sim \rho_\pi}[r_t + \alpha \mathcal{H}(\cdot|\mathbf{x}_t)]$, where α is a temperature parameter, ρ_π denotes the marginal trajectory distribution induced by policy π. MERL model and has proven stable and powerful in low dimensional continuous action tasks, such as games and robotic controls [Haarnoja *et al.*, 2018]. However, when facing complex visual problems such as I2IT, where observations and actions are high-dimensional, it remains a challenge for RL models [Lee *et al.*, 2020].

3 Method

This work reformulates I2IT as a decision making problem. Instead of directly mapping the input to the target image in a single step, we propose to use a light-weight RL model that performs the translation progressively, where new details can be added as the translation progresses. The proposed model *stochastic actor-executor-critic* (SAEC) consists of three components: an actor (π_ϕ), an executor (η_ψ), and a critic (Q_θ). Specifically, the actor π_ϕ and the executor η_ψ form a DL pathway that directs the agent learning with a DL-based objective. The same actor and the critic Q_θ form a DRL pathway that works as an actor-critic model. In SAEC, the actor generates a latent action according to a stochastic policy, so as to capture the required image translation control. The critic evaluates the stochastic policy. The executor leverages the latent action and image details to perform image translation and the generated image is then applied to the environment as the new input image at next time.

3.1 I2IT as MDP

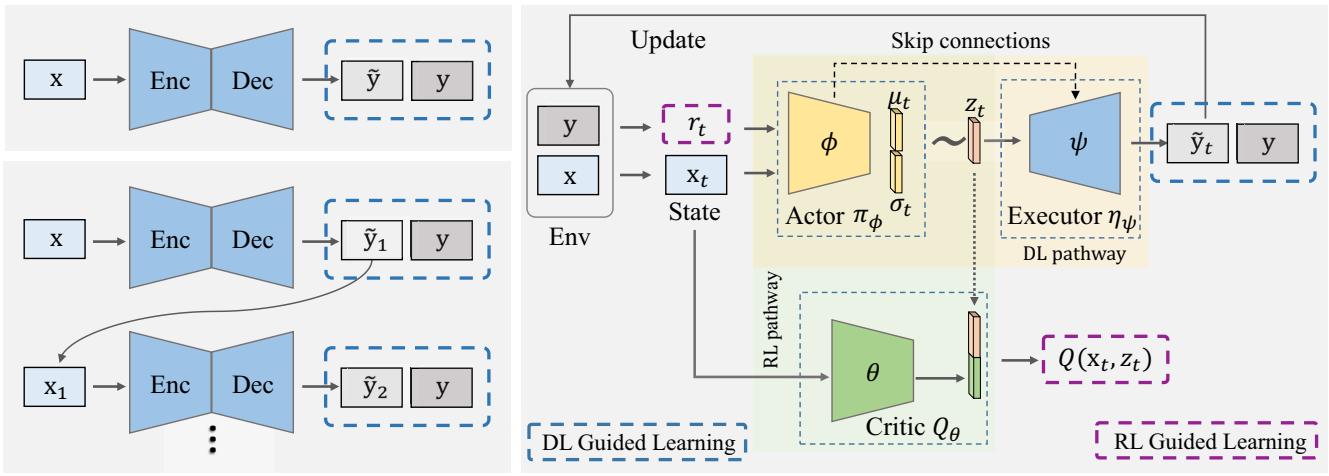
The key of our method is to formulate I2IT as a Markov decision process (MDP). As for state $\mathbf{x}_t \in \mathcal{X}$, it is defined task specific. For realistic photo generation or segmentation, the state $\mathbf{x}_t$ is the pair of the fixed source image I_f and the moving image I_{m_t}: $\mathbf{x}_t = (I_f, I_{m_t})$. A new state is generated by warping the moving image with predicted image action $\tilde{\mathbf{y}}_t$: $\mathbf{x}_{t+1} = (I_f, \tilde{\mathbf{y}}_t \circ I_{m_t})$. For face inpainting, the state is composed of the original image $\mathbf{x}$ and the synthesized missing part $\tilde{\mathbf{y}}_t$. The next state is obtained by adding the new predicted image to the missing area: $\mathbf{x}_{t+1} = (\tilde{\mathbf{y}}_t \odot \mathbf{x}_t)$.

There are two types of actions in our action space: latent action $\mathbf{z}_t \in \mathcal{Z}$ and image action $\tilde{\mathbf{y}}_t \in \mathcal{Y}$. These two actions are intrinsically different, in that $\mathcal{Z}$ has a high-level of abstraction and its distribution is unknown. $\mathbf{z}_t$ is sampled from the stochastic policy of the actor π_ϕ, $\mathbf{z}_t \sim \pi_\phi(\mathbf{x}_t)$, while $\tilde{\mathbf{y}}_t$ is created by the executor based on $\mathbf{z}_t$ and the input image. The reward function is flexible and can be any evaluation metric that measures the similarity between the predicted image $\tilde{\mathbf{y}}_t$ and the target $\mathbf{y}$, such as the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM).

The sequence of actions are implemented by the *stochastic actor-executor-critic* (SAEC) model. Specifically, at each time step t, the actor samples a latent action $\mathbf{z}_t$ from state $\mathbf{x}_t$, then the executor generates image action $\tilde{\mathbf{y}}_t$ from the latent action $\mathbf{z}_t$ and state $\mathbf{x}_t$. The change of the state will result in a reward r_t, and the new state is fed back to the SAEC model to generate a new image action. Let $\mathcal{T}_t$ represents a single run from state $\mathbf{x}_t$ to image action $\tilde{\mathbf{y}}_t$ via an actor and an executor: $\mathbf{z}_t \sim \pi_\phi(\mathbf{x}_t), \tilde{\mathbf{y}}_t = \eta_\psi(\mathbf{x}_t, \mathbf{z}_t)$. The whole process of I2IT in MDP can be formulated as follows:

$$\mathcal{T}(\mathbf{x}) = \mathcal{T}_t \circ \mathcal{T}_{t-1} \circ \cdots \circ \mathcal{T}_0(\mathbf{x}), \tag{1}$$

where $\circ$ is the function composition operation, $\mathcal{T}_t$ is the tth translation step that predicts image from current state $\mathbf{x}_t$.



(a) DL frameworks (b) Stochastic Actor-Executor Critic

Figure 1: The conventional DL frameworks and our SAEC framework. Top of (a): One-stage DL framework. Bottom of (a): Multi-stage DL framework, whose networks are usually different in each stage. (b) Our SAEC framework. $\mathbf{x}$ and $\mathbf{y}$ are input and target, $\tilde{\mathbf{y}}$ is the predicted output. $\mathbf{x_t}$ and $\tilde{\mathbf{y}}_t$ are temporary inputs and outputs generated in step (stage) t. In the SAEC framework, there are two actions: high-level latent-action $\mathbf{z}_t$ which is sampled from policy π_ϕ: $\mathbf{z_t} \sim \pi_\phi(\mathbf{z}_t|\mathbf{x}_t)$, and low-level image-action $\tilde{\mathbf{y}}_t$ which is generated from executor η_ψ conditioned by $\mathbf{z}_t$: $\tilde{\mathbf{y}}_t = \eta_\psi(\tilde{\mathbf{y}}_t|\mathbf{x_t}, \mathbf{z_t})$. Both actions are high-dimensional continuous.

3.2 Policy with Skip Connections

As in an MERL model, the stochastic policy and maximum entropy in our SAEC improve the exploration for more diverse generation possibilities, which helps to prevent agent from producing a single type of plausible output during training (known as mode-collapse). In addition, one specific characteristic in the SAEC model is that skip connections are added from each down-sampling layer of the actor to the corresponding up-sampling layer of the executor, as shown in Fig. 1(b). In this way, a natural looking image is more likely to be reconstructed since the details of state $\mathbf{x}_t$ can be passed to the executor by skip connections. Besides, since both $\mathbf{z}_t$ and $\mathbf{x}_t$ can be used by the executor to generate image action, over-exploration of the action space can be avoided in SAEC, where the variance is limited by these passed detail information. Furthermore, the skip connections also facilitate back-propagation of the DL gradients to the actor. Our experiments show that if the skip connections are removed, the details of the reconstructed image may lost because of down-sampling (max-pooling) and stochastic policies, and the training would be very unstable in such a situation. Combining skip connections and DL-based learning can avoid detail loss and thus stabilize training. (See experiments section for more details).

3.3 Stochastic Actor-Executor-Critic

To address I2IT, We propose to extend the maximum entropy RL (MERL) framework with an additional executor. Our proposed model has a similar structure with SAC [Haarnoja *et al.*, 2018], but the latter cannot be applied to the I2IT tasks directly since the delayed and sparse rewards making it difficult to handle the inherent high-dimensional states and actions. To this end, we propose to train the actor and executor in a conventional DL-based manner, where the training data are collected from the agent-environmental interaction and change dynamically in experience replay. The objective of SAEC is to learn a policy π_ϕ and an executor η_ψ, by maximizing both the conditional likelihood and the expected sum of rewards with an entropy $\mathcal{H}$.

Algorithm 1 Stochastic Actor-Executor Critic

Input: Environment Env and initial parameters θ_1, θ_2 for the Critics, ϕ, ψ for the Actor and Executor.
$\bar{\theta}_1 \leftarrow \theta_1, \bar{\theta}_2 \leftarrow \theta_2, \mathcal{D} \leftarrow \varnothing$
for each iteration **do**
  $\mathbf{x}, \mathbf{y} \sim Env_{reset}()$
  for each environment step **do**
    $\mathbf{z}_t \sim \pi_\phi(\mathbf{z}_t|\mathbf{x}_t)$
    $\tilde{\mathbf{y}}_t \leftarrow \eta_\psi(\mathbf{x}_t, \mathbf{z}_t)$
    $\mathbf{x}_{t+1}, r_t \sim Env_{step}(\tilde{\mathbf{y}}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{y}, \mathbf{x}_t, \mathbf{z}_t, r_t, \mathbf{x}_{t+1})\}$
  for each gradient step **do**
    Update Actor and Executor (DL guided):
      $\phi \leftarrow \phi - \lambda_{DL}\hat{\nabla}_\phi \mathcal{L}_{DL}(\phi)$
      $\psi \leftarrow \psi - \lambda_{DL}\hat{\nabla}_\psi \mathcal{L}_{DL}(\psi)$
    Update Actor and Critic (RL guided):
      $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
      $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi (J_\pi(\phi))$
      $\alpha \leftarrow \alpha - \lambda \hat{\nabla}_\alpha J(\alpha)$
      $\bar{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i$ for $i \in \{1, 2\}$
Output: $\theta_1, \theta_2, \phi, \psi$

Conditional Generative Model Learning

The actor and the executor form a conditional generative process that translates the current state $\mathbf{x}_t$ to target $\mathbf{y}$. Instead of learning the generative model alone like a VAE, we maximize

the conditional likelihood following the stochastic policy π_ϕ:

$$p(\mathbf{y}|\mathbf{x}_t) = \int \pi_\phi(\mathbf{z}_t|\mathbf{x}_t)\eta_\psi(\mathbf{y}|\mathbf{x}_t, \mathbf{z}_t)d\mathbf{x}_t. \quad (2)$$

The empirical objective is to minimize the reconstruction error over all samples:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x}_t,\mathbf{y}\sim\mathcal{D}}[\|\mathcal{T}(\mathbf{x}_t), \mathbf{y}\|_d], \quad (3)$$

where $\mathcal{D}$ is a replay pool, $\|\cdot\|_d$ denotes some distance measures, such as L_1 and L_2. To synthesis more realistic images, we also extend the actor-executor to a high quality generative model by jointly training an adversarial loss with discriminator D:

$$\mathcal{L}_{adv} = \mathbb{E}_{\mathbf{x}_t,\mathbf{y}\sim\mathcal{D}}[\log(D(\mathbf{y})) + \log(1 - D(\mathcal{T}(\mathbf{x}_t)))] \quad (4)$$

Our final DL guided objective can be expressed as

$$\mathcal{L}_{DL} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}, \quad (5)$$

where λ_{rec} and λ_{adv} are used to balance the reconstruction and adversarial learning. In general, the DL pathway is very flexible and can readily leverage any other advanced DL losses or techniques into this framework.

MERL in I2IT

In RL pathway, the rewards and the soft Q values are used to iteratively guide the stochastic policy improve. Moreover, the latent-action $\mathbf{z}_t$, is used to estimate soft state-action value for encouraging high-level policies. Let actor and critic with parameters ϕ and θ be the function approximators for the policy π_ϕ and the soft Q-function Q_θ, respectively. The critic parameters θ are trained to minimize the soft Bellman residual:

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{x}_t,\mathbf{z}_t)\sim\mathcal{D}}\left[\frac{1}{2}(Q_\theta(\mathbf{x}_t, \mathbf{z}_t) - \hat{Q}_{\bar{\theta}}(\mathbf{x}_t, \mathbf{z}_t))^2\right], \quad (6)$$

$$\hat{Q}_{\bar{\theta}}(\mathbf{x}_t, \mathbf{z}_t) = r_t + \gamma\mathbb{E}_{\mathbf{x}_{t+1}\sim P}\left[V_{\bar{\theta}}(\mathbf{x}_{t+1})\right],$$

where $V_{\bar{\theta}}(\mathbf{x}_t)$ is the soft state value function which can be computed by

$$V_{\bar{\theta}}(\mathbf{x}_t) = \mathbb{E}_{\mathbf{z}_t\sim\pi_\phi}\left[Q_{\bar{\theta}}(\mathbf{x}_t, \mathbf{z}_t) - \alpha\log\pi_\phi(\mathbf{z}_t|\mathbf{x}_t)\right]. \quad (7)$$

We use a target network $Q_{\bar{\theta}}$ to stabilize training, whose parameters $\bar{\theta}$ are obtained by an exponentially moving average of parameters of the critic network: $\bar{\theta} \leftarrow \tau\theta + (1 - \tau)\bar{\theta}$. The actor parameters ϕ are optimized to learn the policy towards the exponential of the soft Q-function:

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{x}_t\sim\mathcal{D}}\left[\mathbb{E}_{\mathbf{z}_t\sim\pi_\phi}\left[\alpha\log(\pi_\phi(\mathbf{z}_t|\mathbf{x}_t)) - Q_\theta(\mathbf{x}_t, \mathbf{z}_t)\right]\right] \quad (8)$$

In practice, we use two critics with the same structure but different parameters, (θ_1, θ_2), to mitigate positive bias in policy improvement and accelerate training. Moreover, we automatically tune the temperature hyperparameter α by

$$J(\alpha) = \mathbb{E}_{\mathbf{z}_t\sim\pi_t}[-\alpha\log\pi_t(\mathbf{z}_t|\mathbf{x}_t) - \alpha\bar{\mathcal{H}}], \quad (9)$$

where $\bar{\mathcal{H}}$ represents a constant entropy that equals to the negative of the dimension of the latent action spaces.

The complete algorithm of SAEC is described in Algorithm 1. The training process alternates between learning I2IT from DL pathway and policy control from RL pathway. The joint training of DL and RL ensures a fast convergence since the supervised or unsupervised loss is a "highway" to facilitate back-propagation of the gradients to the actor and the executor. As the latent-action $\mathbf{z}_t$ is concatenated into critic (shown in Figure 1(b)), the RL objective and DL objective can cooperate to train the stochastic actor fast and stable.

4 Experiments

In this section, we demonstrate experimentally the effectiveness and robustness of the SAEC framework on several different tasks of I2IT.

4.1 Face Inpainting

Settings In this experiment, we apply SAEC to the problem of face inpainting, which is to fill in pixels in the central area of a face image with synthesized contents that are semantically consistent with the original face and at the same time visually realistic. Celeba-HQ dataset is used in this study, of which $28,000$ images are used for training and $2,000$ images are used for testing. All images have a missing part of size of 64×64 cropped in the center. We compare the SAEC with several recent face inpainting methods, including CE [Pathak *et al.*, 2016], CA [Yu *et al.*, 2018], PEN [Zeng *et al.*, 2019], PIC [Zheng *et al.*, 2019] and RN [Yu *et al.*, 2020].

The actor-executor for our methods uses the same network structure as the encoder-decoder for CE. The network structures of the discriminator also comes from CE but with a different SNGAN loss. Following the previous work [Pathak *et al.*, 2016; Yu *et al.*, 2018; Zheng *et al.*, 2019], we use PSNR and SSIM as the evaluation metrics.

Results and Analysis As shown in Table 1, by using L_1 and adversarial loss in the DL guided learning, our method achieves the best PSNR and SSIM scores comparing with the existing state-of-the-art methods. And as mentioned in the Method section, the reward function is very flexible for our RL framework. Both the PSNR and SSIM based reward are suitable for SAEC and can improve the performance on face inpainting. The qualitative comparison shown in Figure 2 illustrates that the SAEC gives obvious visual improvement for synthesizing realistic faces. The results of SAEC are very reasonable, and the generated faces are sharper and more natural. This may attribute to the high-level action $\mathbf{z}_t$, which focuses on learning the global semantic structure and then directs the executor with DL guided learning to further improve the local details of the generated image. We can also see that the synthesised images of the SAEC could have very different appearances from the ground truth, which indicates that although our training is based on paired images, the SAEC can successfully explore and exploit data for producing diverse results.

Ablation Study To illustrate the stability of training GANs in SAEC, we jointly use L_1 and several advanced GAN techniques i.e. WGAN-GP [Gulrajani *et al.*, 2017], RaGAN

Figure 2: Visual comparison of face inpainting over all methods. Our SAEC use SNGAN for auxiliary DL guided learning. # indicates what reward is used for RL training. Our results have better visual quality even for the large pose face.

Method	PSNR ↑	SSIM ↑
CE [Pathak *et al.*, 2016]	26.949	0.870
CA [Yu *et al.*, 2018]	26.608	0.871
PIC [Zheng *et al.*, 2019]	26.307	0.893
PEN [Zeng *et al.*, 2019]	26.391	0.883
RN [Yu *et al.*, 2020]	26.932	0.892
SAEC (SNGAN + PSNR reward)	27.176	0.882
SAEC (SNGAN + SSIM reward)	**27.327**	**0.896**

Table 1: Quantitative results of all methods on Celeba-HQ.

Method	PSNR ↑	SSIM ↑
AE + SNGAN	26.884	0.871
SAEC (+ WGAN-GP)	27.091	0.875
SAEC (+ RaGAN)	27.080	0.873
SAEC (+ SNGAN)	**27.176**	**0.882**

Table 2: Quantitative results of the variants methods on Celeba-HQ testing dataset (all trained with PSNR reward).

[Jolicoeur-Martineau, 2018], and SNGAN [Miyato *et al.*, 2018] for DL guided learning. To validate the effectiveness of RL pathway in SAEC for I2IT, we also separately train an actor-executor model (AE) by jointly optimizing the L_1 and SNGAN loss. The results shown in Table 2 indicate that the proposed SAEC with different GANs are stable, and significantly improve the performance of training AE with SNGAN alone, which further demonstrates the contribution of the RL formulation.

4.2 Realistic Photo Translation

Settings In this section, we evaluate the SAEC on a set of general I2ITs. We select three datasets of realistic photo translation including:

- CMP Facades dataset for segmentation *labels→images* [Tyleček and Šára, 2013].
- Cityscapes dataset for segmentation *labels→images* and *images→labels* [Cordts *et al.*, 2016].
- Edge and shoes dataset for *edges→shoes* [Yu and Grauman, 2014].

In these tasks, SAEC uses the same network structure as in face inpainting experiment with PSNR reward and SNGAN loss. We compare our method with pix2pix [Isola *et al.*, 2017] and PAN [Wang *et al.*, 2018a]. We also compared with pix2pixHD [Wang *et al.*, 2018b], DRPAN [Wang *et al.*, 2019] and CHAN [Gao *et al.*, 2021], which are designed for high-quality I2IT. Moreover, we replace MERL with PPO [Schulman *et al.*, 2017] in the RL pathway as SAEC*. We use PSNR, SSIM and LPIPS [Zhang *et al.*, 2018] as the evaluation metrics.

Quantitative Evaluation The quantitative results are shown in Table 3. With a similar network structure, the proposed method significantly outperforms the pix2pix and PAN model on PSNR, SSIM and LPIPS over all datasets and tasks. The SAEC even achieves a comparable or better performance than the high-quality pix2pixHD and DRPAN model, which have much more complex architectures and training strategies. Moreover, using MERL instead of PPO obviously improves performance on most tasks. These experiments illustrate that the proposed SAEC is a robust and effective solution for I2IT.

Qualitative Evaluation The qualitative results of our SAEC with other I2IT methods on different tasks are shown in Figure 3. We observe that the pix2pix and PAN sometimes suffer from mode collapse and yield blurry outputs. The pix2pixHD is unstable in different datasets especially on Facades and Cityscapes. The DRPAN is more likely to produce blurred artifacts in several parts of the predicted image on Cityscapes. In contrast, the SAEC produces more stable and realistic results. Using stochastic policy and MERL helps to explore more possible solutions so as to seek out the best generation strategy by trial-and-error in the training steps, leading to a more robust agent for different datasets and tasks.

Comparison with other RL Algorithms The key components of SAEC are substituted by other structures or other state-of-the-art RL algorithms to test their importance. We use DDPG and PPO to demonstrate the effectiveness of stochastic policy and maximum entropy RL, respectively. The learning curves of different variants on four tasks are shown in Figure 4. By using the stochastic policy and the maximum entropy framework, the training is significantly improved.

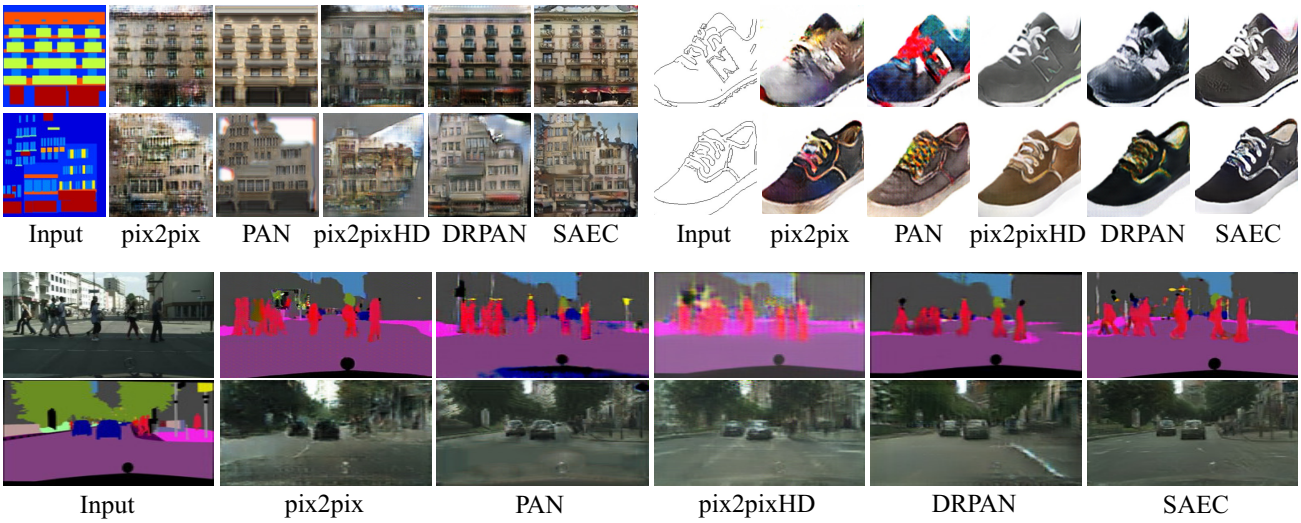


Figure 3: Visual comparison of our method with pix2pix, PAN, pix2pixHD and DRPAN over all tasks.

Method	Facades label→image			Cityscapes image→label			Cityscapes label→image			Edges→shoes		
	PSNR	SSIM	LPIPS ↓	PSNR	SSIM	LPIPS ↓	PSNR	SSIM	LPIPS ↓	PSNR	SSIM	LPIPS ↓
pix2pix	12.290	0.225	0.438	15.891	0.457	0.287	15.193	0.279	0.379	15.812	0.625	0.279
PAN	12.779	0.249	0.387	16.317	0.566	0.228	16.408	0.391	0.346	16.097	0.658	0.228
pix2pixHD	12.357	0.162	0.336	17.606	0.581	0.204	15.619	0.361	**0.319**	17.110	0.686	0.220
DRPAN	13.101	0.276	0.354	17.724	0.633	0.214	16.673	0.403	0.343	17.524	**0.713**	0.221
CHAN	13.137	0.231	0.402	17.459	0.641	0.222	16.739	0.401	0.373	18.065	0.692	0.236
SAEC*	13.163	**0.308**	0.366	17.168	0.616	0.221	16.685	0.410	0.362	16.914	0.695	0.225
SAEC	**13.178**	0.296	**0.324**	**17.969**	**0.659**	**0.203**	**16.848**	**0.412**	0.337	**18.178**	0.698	**0.215**

Table 3: Quantitative results of our SAEC with other methods over all datasets. ↓ means lower is better, SAEC* means using PPO.

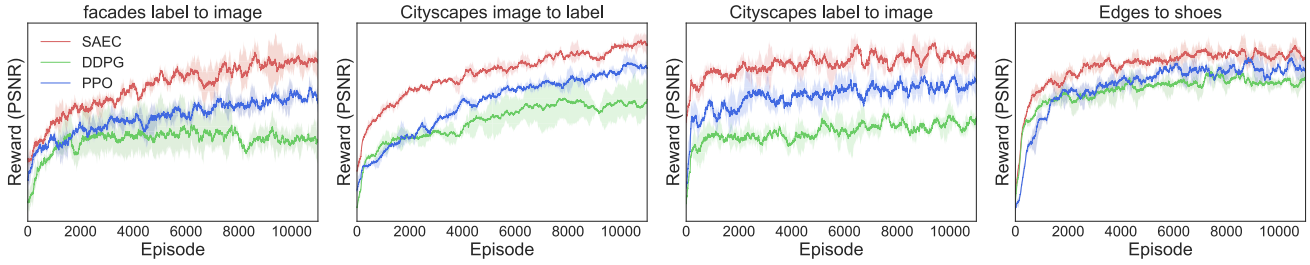


Figure 4: Learning curves on different I2IT tasks. SAEC performs consistently better than other modified RL algorithms.

5 Conclusion

In this paper, we present a new RL framework for high-dimensional continuous control and apply it to I2IT problems. This framework contains a *stochastic actor-executor critic* (SAEC) structure that handles high dimensional policy control, high-level state representation, and low-level image generation simultaneously. We train SAEC by jointly optimizing the RL and the DL objective, which provides ubiquitous and instantaneous supervision for learning I2IT. Our empirical results on several different vision applications demonstrate the effectiveness and robustness of the proposed SAEC. We will subsequently extend this work to more computer vision tasks involving with high-dimensional continuous spaces.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61602065, Sichuan province Key Technology Research and Development project under Grant 2021YFG0038.

References

[Caicedo and Lazebnik, 2015] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2488–2496, 2015.

[Cordts *et al.*, 2016] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Gao *et al.*, 2021] Fei Gao, Xingxin Xu, Jun Yu, Meimei Shang, Xiang Li, and Dacheng Tao. Complementary, heterogeneous and adversarial networks for image-to-image translation. *IEEE Transactions on Image Processing*, 30:3487–3498, 2021.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[Isola *et al.*, 2017] 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.

[Jolicoeur-Martineau, 2018] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[Lee *et al.*, 2020] Alex X. Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[Miyato *et al.*, 2018] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[Nair *et al.*, 2018] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, pages 9191–9200, 2018.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Tyleček and Šára, 2013] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, pages 364–374. Springer, 2013.

[Wang *et al.*, 2018a] Chaoyue Wang, Chang Xu, Chaohui Wang, and Dacheng Tao. Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, 27(8):4066–4079, 2018.

[Wang *et al.*, 2018b] 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.

[Wang *et al.*, 2019] Chao Wang, Wenjie Niu, Yufeng Jiang, Haiyong Zheng, Zhibin Yu, Zhaorui Gu, and Bing Zheng. Discriminative region proposal adversarial network for high-quality image-to-image translation. *International Journal of Computer Vision*, 2019.

[Yarats *et al.*, 2019] Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. *arXiv preprint arXiv:1910.01741*, 2019.

[Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[Yu *et al.*, 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[Yu *et al.*, 2020] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region normalization for image inpainting. In *AAAI*, pages 12733–12740, 2020.

[Zeng *et al.*, 2019] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1486–1494, 2019.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zheng *et al.*, 2019] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.