Unsupervised Image-to-Image Translation with Stacked Cycle-Consistent Adversarial Networks

Minjun Li[1,2], Haozhi Huang[2], Lin Ma[2], Wei Liu[2], Tong Zhang[2], Yu-Gang Jiang[1]

[1]School of Computer Science, Fudan University, [2]Tencent AI Lab me@minjun.li, {huanghz08, forest.linma}@gmail.com, wl2223@columbia.edu, tongzhang@tongzhang-ml.org, ygj@fudan.edu.cn

1. Motivation

Recent studies on unsupervised image-to-image translation have made remarkable progress by training generative adversarial networks with the cycle-consistent loss. However, such methods still have the following potential problems:

1. May generate inferior results, especially when the image resolution is high.
2. Unable to learn reasonable translations when the two image domains are of significant differences.

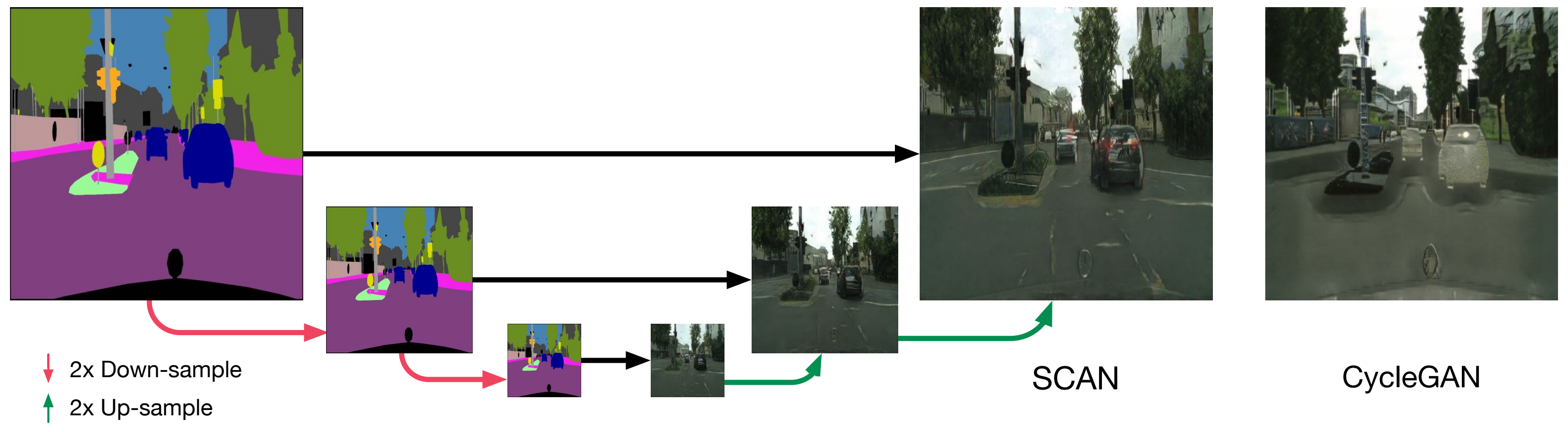
2. Contribution

We propose the Stacked Cycle-Consistent Adversarial Networks (namely SCAN) by using a coarse-to-fine approach, which not only boosts the image translation quality but also enables higher resolution and more difficult translation.

1. The SCAN models unsupervised image-to-image translation problem in a coarse-to-fine manner, which generates finer details in higher resolution and enables learning of more difficult image transitions.
2. The adaptive fusion block dynamically integrates output from different stages, which outperforms the uniform-weight fusion approaches.

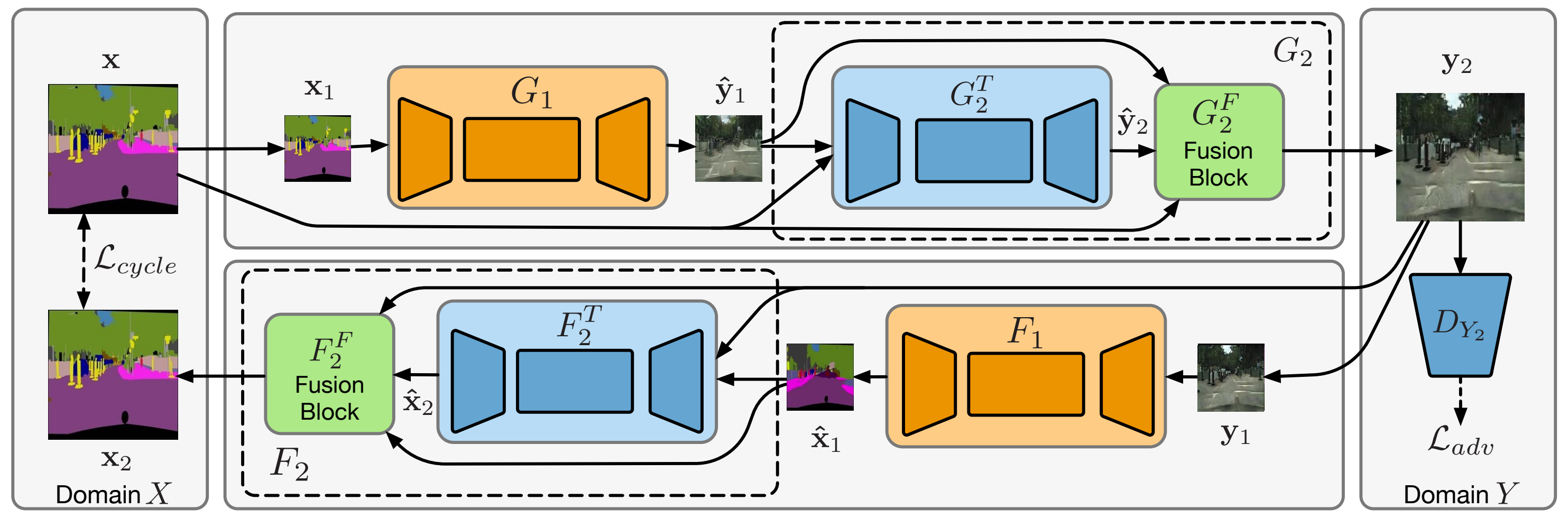
3. Architecture

Given unpaired images from two domains, our proposed SCAN learns the image-to-image translation by a stacked structure in a coarse-to-fine manner. For the Cityscapes *Labels* → *Photo* task in 512 × 512 resolution, the result of SCAN appears more realistic and includes finer details compared with the result of CycleGAN [Zhu *et al.*2017].

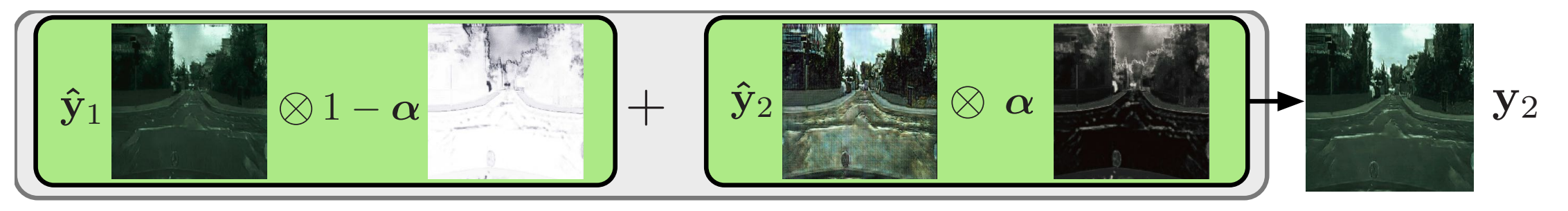


After the coarse translations are learned, SCAN learns the refining processes on the top of previous stage's outputs. In the training process, we keep the weights of previous stages fixed.

3. Architecture (Cont.)

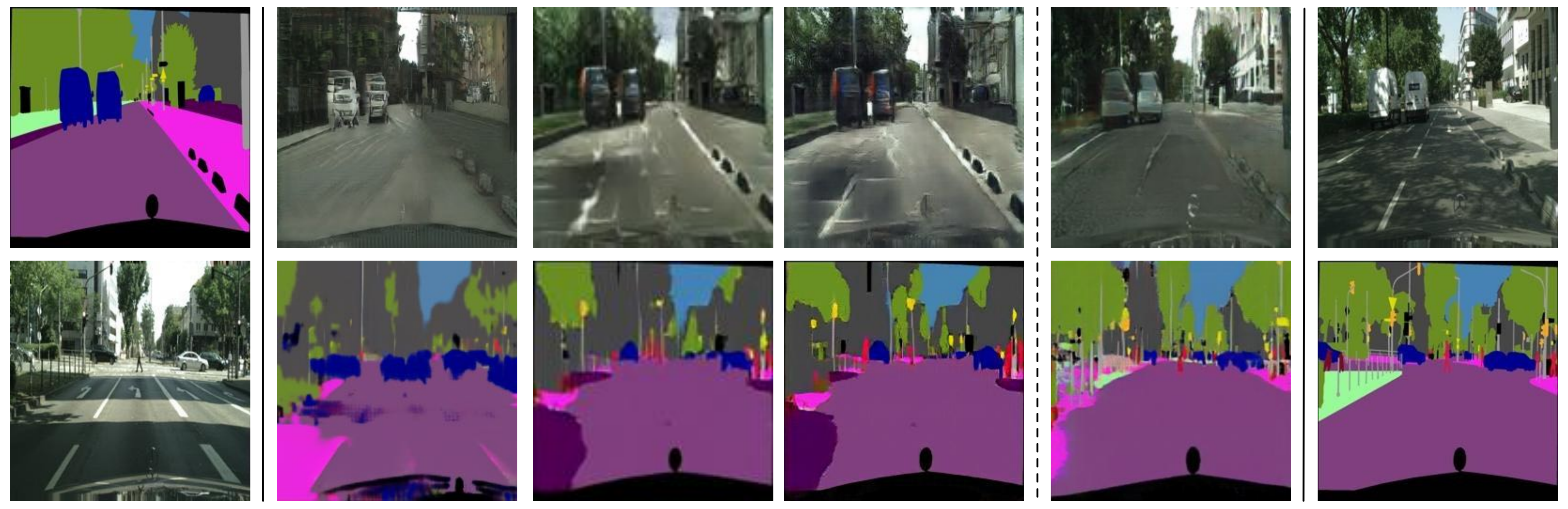


The fusion block applies the fusion weight map to find defects in the previous results and correct it to produce refined output.



4. Experiments

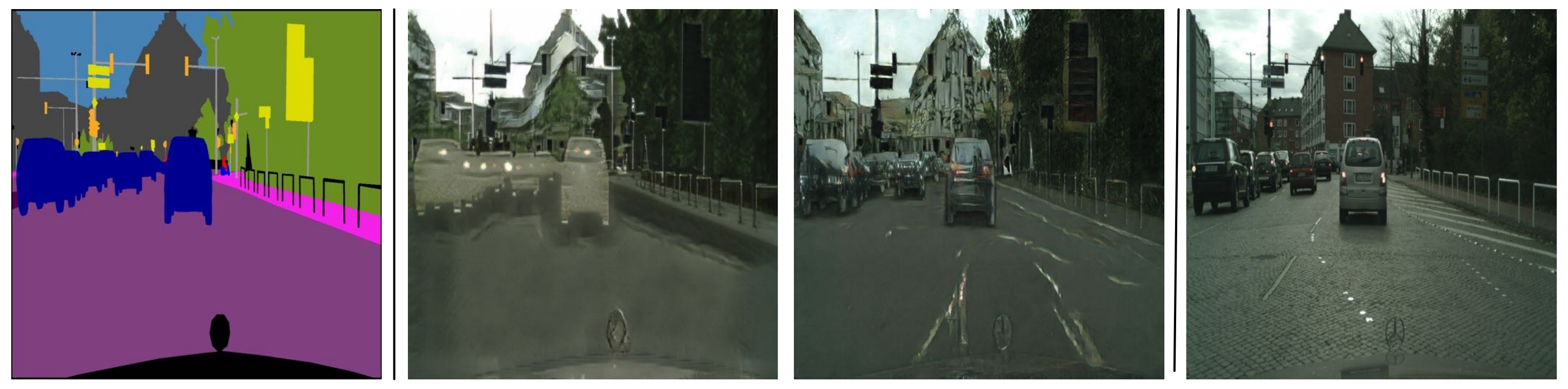
Comparisons on the Cityscapes dataset in 256 × 256 resolution. SCAN generates more natural photographs and accurate segmentations than CycleGAN, and appears closer to supervised approach (Pix2Pix [Isola *et al.*2017]).



Input CycleGAN SCAN Stage-1 SCAN Stage-2 Pix2Pix Ground Truth

	Labels → Photo			Photo → Labels		
Method	Pixel acc.	Class acc.	Class IoU	Pixel acc.	Class acc.	Class IoU
CycleGAN [Zhu *et al.* 2017]	0.52	0.17	0.11	0.58	0.22	0.16
Contrast-GAN [Liang *et al.* 2017]	0.58	**0.21**	**0.16**	0.61	0.23	0.18
SCAN Stage-1 128	0.46	0.19	0.12	0.71	0.24	0.20
SCAN Stage-1 256	0.57	0.15	0.11	0.63	0.18	0.14
SCAN Stage-2 256-256	0.52	0.15	0.11	0.64	0.18	0.14
SCAN Stage-2 128-256	**0.64**	0.20	**0.16**	**0.72**	**0.25**	**0.20**
Pix2Pix [Isola *et al.* 2017]	0.71	0.25	0.18	0.85	0.40	0.32

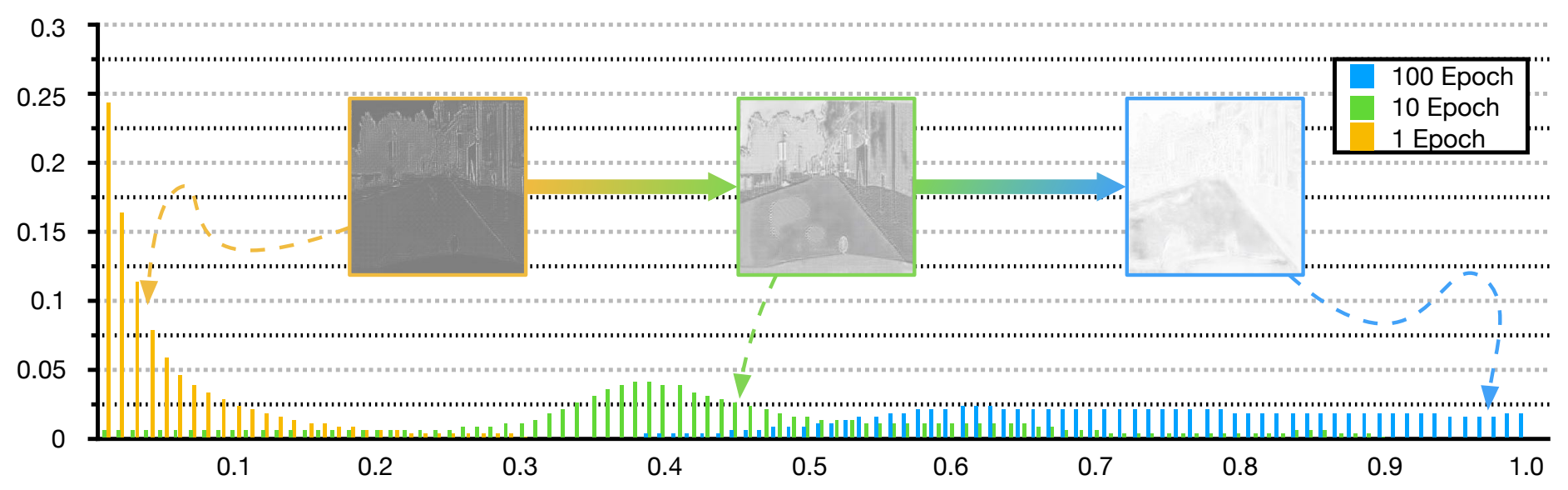
SCAN also enables learning higher resolution translations under unsupervised setting. *e.g.Labels* → *Photo* in 512×512 resolution.



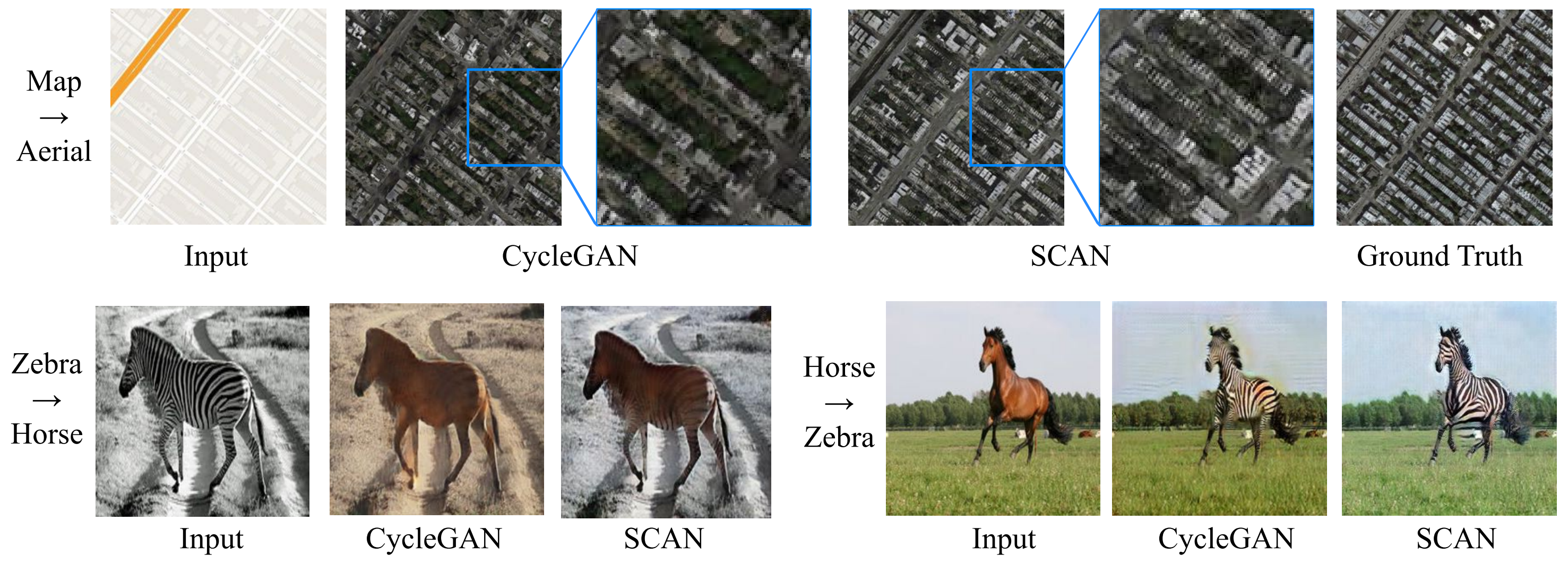
Input CycleGAN SCAN Ground Truth

4. Experiments (Cont.)

The average weights of the fusion maps $\boldsymbol{\alpha}$ increase consistently during training, which implies more and more details of the second stage are brought to the final output.

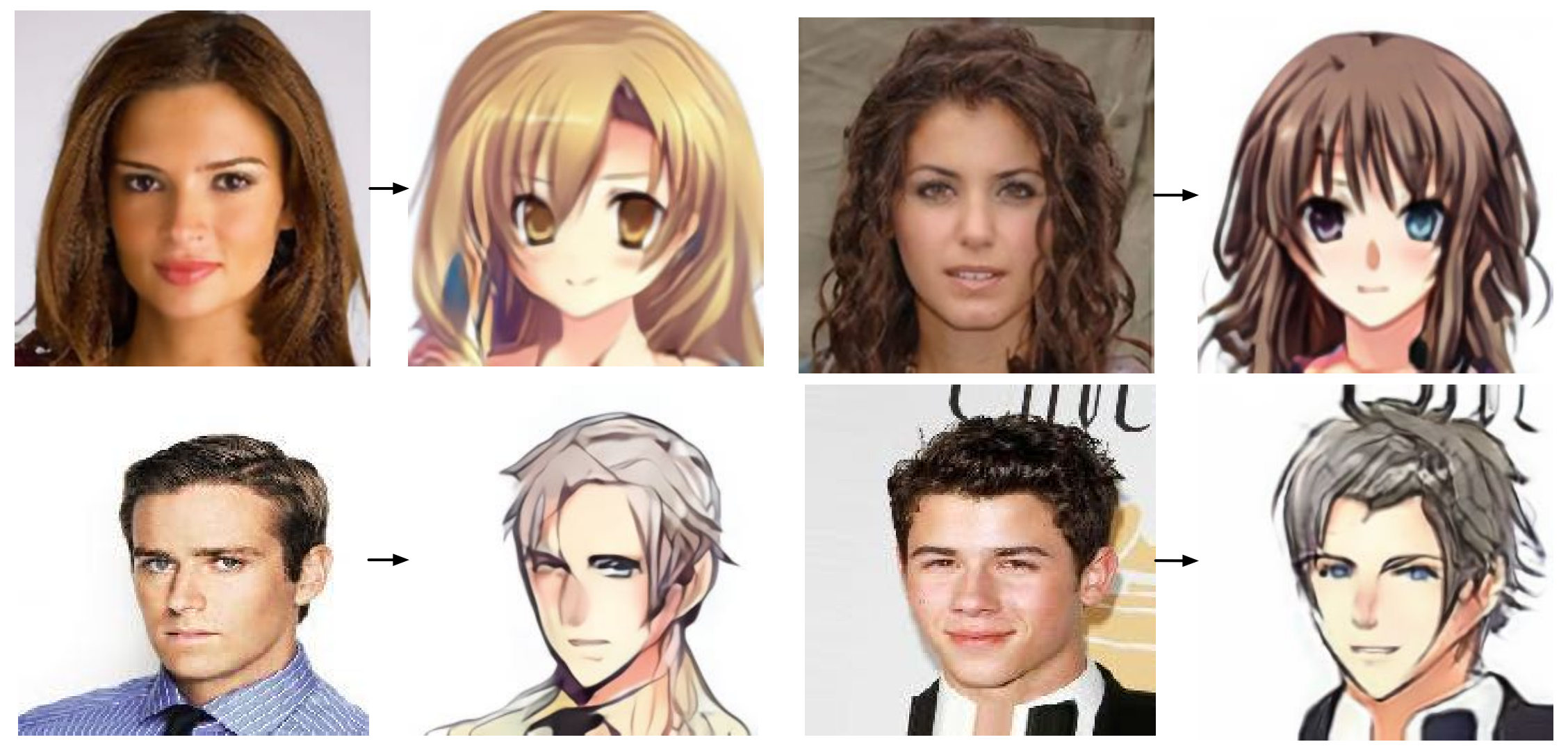


Quality of other image translation tasks can also be improved regarding image sharpness and overall consistency.



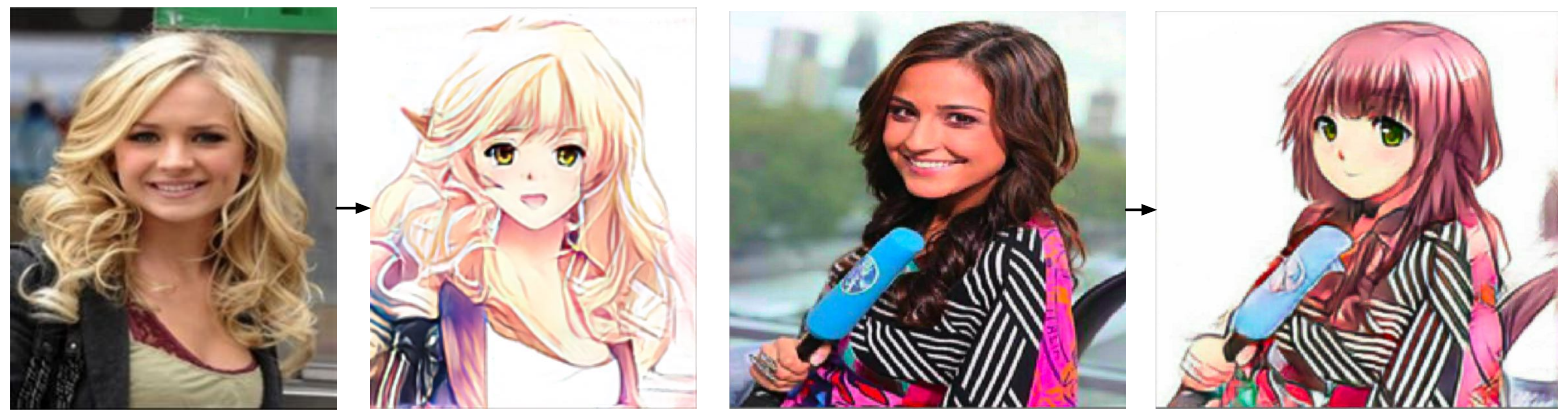
5. Applications (Human2Anime)

Using SCAN, we can learn difficult translation even when paired training data is unavailable, *e.g.* translate real person head portrait into anime avatars.



*Human images are from CelebA dataset [Liu *et al.*2015].

Since we use fully convolutional structure, the network learned from head portraits can also directly apply to various photographs.



*Human images are from CelebA dataset.