Perceptual Adversarial Networks With A Feature Pyramid for Image Translation

Zhuorong Li
Zhejiang University City College

Minghui Wu
Zhejiang University City College

Corresponding Author

Jianwei Zheng
Zhejiang University of Technology

Hongchuan Yu
Bournemouth University

This work investigates the image-to-image translations problems, where the input image is translated into its synthetic form with the original structure and semantics preserved. Widely used methods compute the pixel-wise MSE loss, which are often inadequate for high-frequency content and tend to produce overly smooth results. Concurrent works that leverage recent advances in conditional generative adversarial networks (cGANs) are proposed to enable a universal approach to diverse image translation tasks that traditionally require specific loss functions. Despite the impressive results, most of these approaches are notoriously unstable to train and tend to induce blurs. In this paper, we decompose the image into a set of images by a feature pyramid and elaborate separate loss components for images of specific band-pass. The overall perceptual adversarial loss is able to capture not only the semantic features but also the appearance.

Many classical image processing tasks can be framed as image-to-image translation problems, where an input image would be mapped to a synthetic counterpart.[1] Examples include cross-domain image generation, image colorization, image de-nosing, image inpainting, image semantic segmentation, image super-resolution and texture transfer, *etc*. Such image-to-image translation is useful in generating synthetic images for many downstream tasks. For instance, the translated images can be used for supplementing the missing data or producing considerable sizes of data for regression or classification tasks.

The surge of interest in convolutional neural networks has exerted significant impact on image translation tasks. Subsequently training a convolutional neural network or its variant is a prominent approach for image processing. However, the loss functions used by these methods are all based on pixel-wise construction errors between the predicted image and the ground-truth, which are notoriously inadequate for high-level representations and will tend to yield blurry reconstructions.[2] This is because the minimization of pixel-wise loss encourages the averaging of all plausible locations rather than preserves the precise location of details in the features.

GANs have been recently proposed as a novel approach to train a generative model,[3] providing an appealing alternative to image generation without resorting to detailed goals. We give a brief review of the closely related work of this paper by focusing on two areas: effective optimization objectives and efficient network architectures. Zhu *et al* [4] and Isola *et al* [5] move beyond specific image translation tasks by developing a cGAN-based common framework [6] for various image-to-image translation tasks. Taigman *et al* [7] extend to the unsupervised version, mapping images from source domain to target domain instead of specific input and output images. Unfortunately, GANs are notoriously unstable to train in practice. Recent papers have shown that GAN-based methods that integrate the perceptual loss, which are based on difference in high-level feature space, is able to yield pleasing synthetic images and also help stabling training. [8,9] For tasks that allow for nearly optimal solutions, like texture transfer,[9] desirable outputs can be attained by matching the perceptual features. However, minimizing the perceptual distance alone is not able to tackle more complicated problems like extreme super-resolution [2] and also tends to induce high-frequency artifacts.

Another intriguing line of related work consists of assembled GANs architectures. CycleGAN,[4] DualGAN [10] and DiscoGAN [11] slove the unpaired image translation by enforcing a cyclic loss between the source domain and the target domain, which might compete more with the adversarial losses especially in the paired image translation tasks. In contrast, CoGAN [12] does not rely on the invertibility of the bi-directional mapping, instead it learns the joint distribution by a weight-sharing assumption. One major criticism has been its poor universality since the joint representation across domains is task-specific. Wang *et al* [13] and Chen *et al* [14] exploit the coarse-to-fine generator. Inspired by their successes, we propose a new successive refinements scheme.

Our work resembles in spirit what Wang *et al* [13] have done in terms of the multi-resolution pipeline, but technically is very different. With respect to the discriminators, rather than training the critics to differentiate the synthesized images and the real ones in the image space, we utilize the discriminators to minimize the discrepancy in dual-track feature space at multiple scales. As for the generators, downsides of Wang *et al* [13] are large memory for concatenating input into the generator and the inapplicability to scenes that inaccessible to the manually annotated labels for each individual object.

In this paper, a cGAN-based framework is developed for the image-to-image translation that would traditionally require tailored loss function with expert knowledge. We introduce a feature pyramid to separate the high-frequency and low-frequency feature maps thus effectively alleviate either the problem of high-frequency artifacts or over smooth in image translation. Besides, in order to capture the perceptually important information, deep hierarchical discrepancy is incorporated into the adversarial training. We also conduct an ablation study regarding the optimizing objectives and the essentiality of the integrated feature pyramid. Comparisons against baselines on several image translation tasks have been performed and the method proposed is demonstrated to be effective.

METHODS

The goal of image translation is to learn a mapping from images in domain X to those in Y given training set $\{x_i\}_{i=1}^{m} \in X$ and $\{y_j\}_{j=1}^{n} \in Y$. In this paper we focus on the paired image-to-image translation, where each y_j is matched with x_i accordingly. As can been illustrated in Figure 1, the input image is translated into its counterpart that with the characteristics of the target domain and then decomposed into two sets of images by feature pyramid structures. High-

frequency versions of the input image and the output image are learned for semantics comparison. On the other hands, low-frequency residual images are used for spatial resolution to ensure the global coordination. Besides, feature matching in pixel-wise space as well as perceptual distance are adopted to provide gradients that alleviate the unstable training problem of GANs to some extent.

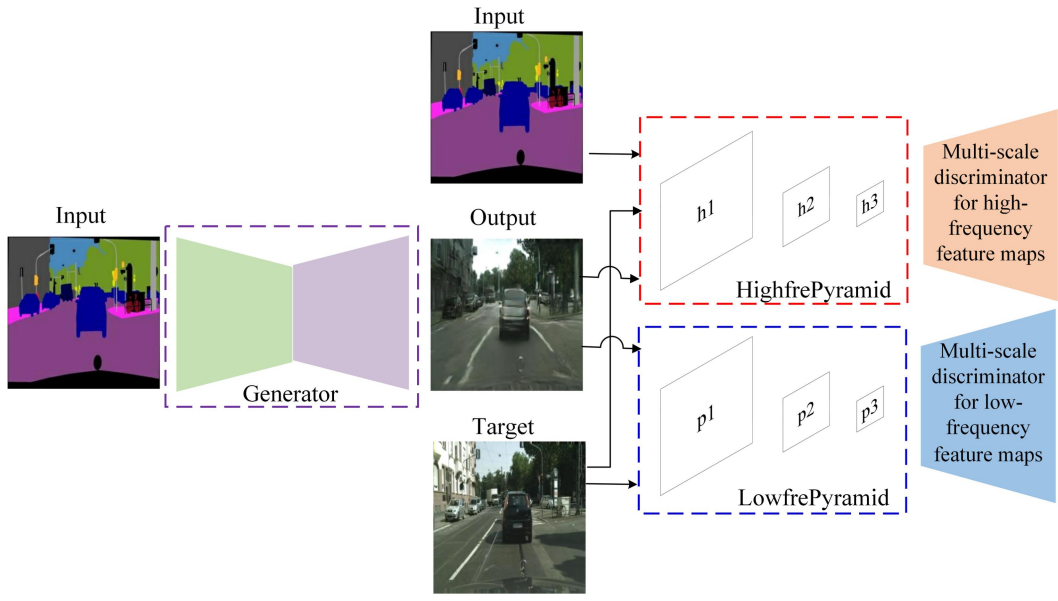


Figure 1: Schematic diagram of the proposed algorithm with illustrative images.

Our ultimate aim is to translate for a given image its corresponding counterpart with high perceptual quality. To achieve this, we train a generator G by solving:

$$\hat{\theta}_G = \arg\min_{\theta_G} \sum_i \alpha_i \mathrm{L}_i \left(G_{\theta_G}(X), Y \right) \qquad (1)$$

where θ_G is the parameters of the generator G and the summation denotes the weighted combination of several loss functions that depict distinct characteristics of natural images.

We utilize a feature pyramid structure [15] to obtain image representations that of different band-pass. Specifically, a image I of size $w \times h$ is downsampled to a new one with the size of $w/2 \times h/2$ by the operator $d(.)$ and get blurred. Keep downsampling we get a pyramid $P(I)=[p_0(I), p_1(I),\ldots,p_k(I)]$ that consist of a set of images with various scales, i.e. $p_2(I)=d(d(I))$ with the size of $w/4 \times h/4$. Each level of the high-frequency pyramid $H(I)==[h_0(I), h_1(I),\ldots,h_k(I)]$ is defined by the residual image between forwarding adjacent level in the pyramid $P(I)$. Note that the level with smaller scale should be upsampled to ensure the element-wise subtracting. Formally, the k-th high-frequency residual image of pyramid is:

$$h_k = p_k - u\left(p_{k+1}\right) \qquad (2)$$

where $u(.)$ is an upsampling operator that doubles the image scale. Constructed by first downsampling and then upsampling, $u(p_k+1)$ acts as a low-frequency version of the image.

Then we elaborate separate loss for different feature pyramid images to capture characteristics of both semantics and appearance. GANs [3,6] provide a novel approach to learn a prior. Specifically, A GAN contains two components: a generator to generate realistic fooling images and a discriminator to distinguish the fake from the real image. GANs are just right for general image-to-image translation tasks as they would reject the undesirable outputs no matter what the

specific application is. GANs are satisfying solutions for a wide range of image translation tasks since they can learn a loss that adapts to the data. Formally, the adversarial loss can be written as:

$$L_{GAN} = E_{x\sim p_{data}(x)}\left[\log D(x)\right] + E_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right] \quad (3)$$

where x is the observed image and z is the random noise vector that mapped to the output image by the generator G. In practice, both the discriminator D and the generator G are iteratively optimized by stochastic gradient decent (SGD).

Previous works have shown that performance can be improved by modifying the generator G or the discriminator D. To be more specific, when the up-scaling generator in the naive GAN [3] is replaced by a network architecture that akin to an autoencoder, the input image x is mapped to a latent representation and then produces a fake image $G(x)$ that being expected to have the same distribution with that of the target image y: [5,16]

$$L_{GAN\_AE} = E_{y\sim p_{data}(y)}\left[\log D(y)\right] + E_{x\sim p_{data}(x)}\left[\log\left(1-D(G(x))\right)\right] \quad (4)$$

However, randomness is necessary if deterministic results are expected to avoid. Therefore, dropout is used as noise to produce stochasticity at several layers of the generator G. When it comes to the discriminator, previous works [5,6] have found it beneficial to simultaneously use the input image as the condition variable of discriminator D:

$$L_{cGAN} = E_{x,y\sim p_{data}(x,y)}\left[\log D(x,y)\right] + E_{x\sim p_{data}(x)}\left[\log\left(1-D(x,G(x))\right)\right] \quad (5)$$

Since impressed outputs are always globally consistent with the input image and share the details of appearance with the target images. Therefore we propose to utilize the levels in separate feature pyramid of real image x in the source domain and y in the target domain. To be more specific, we adversarially train the generator G and discriminator D on high-frequency feature maps of the output image $G(x)$ that constructed by a pyramid network to match the semantic characteristics with the input x. We utilize the multi-scale discriminators [13] to tackle pyramid levels that of different sizes. Formally the adversarial loss on high-frequency images can be written as:

$$L_{cGAN}^{high} = \sum_k \left[E_{x,y\sim p_{data}(x,y)}\left[\log D_k\left(P_k(x), h_k(y)\right)\right] + E_{x\sim p_{data}(x)}\left[\log\left(1-D_k\left(P_k(x), h_k(G(x))\right)\right)\right] \right] \quad (6)$$

where $P_k(x)$ is the downsampled image that of $1/k$ scale of the input x and h_k is the k-th level in the high-frequency feature pyramid. D_k denotes the corresponding discriminator with respect to the image scale. Note that though different D_ks operate at distinct image scale, they have identical network architecture.

On the other hand, we train on smooth images that spaced by an octave to capture the global information such as the color and contents coordination of the target y:

$$L_{cGAN}^{low} = \sum_k \left[E_{y\sim p_{data}(y)}\left[\log D_k\left(P_k(y)\right)\right] + E_{x\sim p_{data}(x)}\left[\log\left(1-D_k\left(P_k\left(G(x)\right)\right)\right)\right] \right] \quad (7)$$

As demonstrated in the report by Mahendran and Vedaldi,[17] minimizing the difference in the feature space instead of the image space tends to generate outputs that are undistinguished from the targets. This can be attributed to the ability of deeper layers to represent much complex features than from pixel space. Inspired by Johnson *et al* [1] and Gatys *et al* [8], we introduce a perceptual loss that defined as the Euclidean distance of l-th ReLU activation layers of the perceptual network ϕ:

$$\mathrm{L}_{perceptual} = \sum_l \frac{\beta_l}{C_l H_l W_l} \left(\phi_l \left(G(x) \right) - \phi_l \left(y \right) \right)^2 \quad (8)$$

where C_l, H_l and W_l are filter amount, height and weight respectively, and the weighting parameter β_l describes the contribution of layer l to the sum of perceptual loss. The perceptual discrepancy provides considerable gradients for the generator to be optimized thus encourages the perceptual similarity between the translated image and the reality, rather than just forcing them to be exactly the same in pixel values. The perceptual network ϕ might be the pre-trained deep neural networks such as VGG19, a part of the discriminator or the generator network.[2] Our method uses all the ReLU activation layers of the conditional discriminator network D since preliminary experiments show that the higher layers of our discriminator network are sufficient to provide perceptual representations at an insignificant degradation of visually performance when compared to deeper networks such as VGG19.

However, using perceptual loss alone generally induces high-frequency artifacts.[17] Thus mixing the perceptual loss with other losses is essential to guide the image generation. Despite blurry outputs, the traditional loss like $\ell 1$ or $\ell 2$ distance has been demonstrated to be beneficial when combined with the adversarial loss. [2,13] This is because the pixel loss could provide gradients that alleviate the unstable training problem of GANs to some extent. In this paper, $\ell 1$ distance is used for the pixel-wise loss to encourage less blur than $\ell 2$ loss:

$$\mathrm{L}_{pix} = \mathrm{E}_{x,y \sim p_{data}(x,y)} \left[\left\| G(x) - y \right\|_1 \right] \quad (9)$$

Finally, the full objective of the perceptual adversarial loss we aim to optimize is:

$$\hat{\theta}_G = \arg\min_{\theta_G} \left(\alpha_1 \mathrm{L}^{high}_{cGAN} + \alpha_2 \mathrm{L}^{low}_{cGAN} + \alpha_3 \mathrm{L}_{perceptual} + \alpha_4 \mathrm{L}_{pix} \right) \quad (10)$$

Usually there is no perfect solution that simultaneously achieves the minimum of each loss component. However, since the total loss we optimize is a linear combination of each separate loss function, we can tune the coefficients of them to balance between high-frequency and low-frequency feature learning, as well as place the emphasis on pixel-wise or perceptual similarity. We test for different coefficient combinations in the preliminary experiments to make our model equip with comprehensive capacity on image-to-image transformation tasks. In this paper, we set α_1=1, α_2=5, α_3=1, and α_4=100 respectively.

EXPERIMENTS

We perform an ablation study regarding to the optimizing objectives and the integrated feature pyramid to empirically demonstrate that the proposed method is an effective approach to achieve substantial performance boost. To further assess the proposed approach, we apply it to several image-to-image translation tasks and compare it with baseline models.

Implementation Details

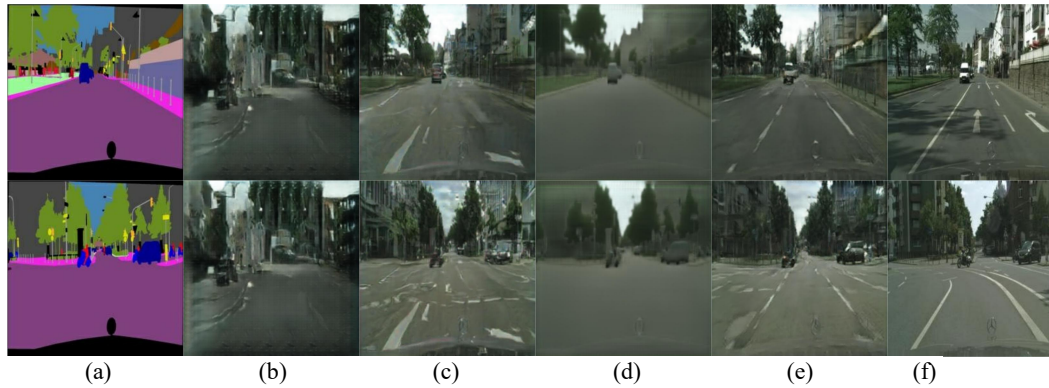
Network architectures in this paper adopt the ones in pix2pix.[5] Besides, a perceptual network ϕ is employed to measures the discrepancy between the generated image and the real ones using the perceptual similarity metric. In our work the perceptual network ϕ shares the same architecture with the discriminator D. As for the discriminator, 70×70 patch-level discriminator is used on purpose of fewer parameters. We apply instance normalization for generators. When it comes to the details of the feature pyramid, preliminary experiments show that three levels for each are enough to yield satisfactory results. Slight improvement by deepening the feature pyramid could not compensate for the computation overhead since the multi-scale discriminators are required to judge the representations at all the levels. Nearest neighbor interpolation and the bicubic interpolation are adopted for downsampling and upsampling respectively.

Networks are trained from scratch with weights initialized from a Gaussian distribution N (0, 0.02^2). Alternate SGD and Adam solver with a momentum term of 0.5 and a learning rate of 2×10^{-4} are applied to *D* and *G*. Training epochs vary with the dataset size of different tasks. For all the implementations, Tensorflow and cuDNN have been employed on a NVIDIA 1080 Ti GPU.

Ablation Study

The significance of each loss component of the overall perceptual adversarial loss is need to be validated by comparing the unconditional version against those conditioned, as well as the partial model against the full model regarding the feature pyramid.

Figure 2 shows the qualitative performance of these variations on the widely used benchmark dataset Cityscapes.[18] When the conditioning of Equation 6 is removed (see Figure 2(a)), the generator produces nearly the same transformed image regardless of different inputs. When the discriminator is conditioned on the input, output varies with the input image rather than just being constant as they are constrained to be matched (see Figure 2(c)-(e)). In other words, mode collapse can be alleviated by conditioning effectively. But still discrepancy between the results of depending on single pyramid and targets can be observed. We found obvious distortion of the zebra crossing in Figure 2(c) and over smoothness in Figure 2(d). Our conjecture is that some divergences exist such as the appearance between the output of the generator and the conditional image, the former being photograph whereas the latter being semantic labels. Then attempt has also been made to leverage both the high-frequency and low-frequency pyramids, which we find really effective at yielding realistic synthesis. Besides, since the discriminator also acts as the perceptual network in our approach, the perceptual discrepancy between the output of generator and the input image brings extra gradients to the generator to be optimized.



(a) (b) (c) (d) (e) (f)

Figure 2: Visually evaluation depending on whether or not to employ the conditioning and what kind of maps in feature pyramid to leverage for Cityscapes labels→photo translation task. From left to right: (a) input; (b) without conditioning; (c) without Equation 7; (d) without Equation 6; (e) completed model; and (f) target.

To further validate the proposed method, the FCN score, [5] which measures the interpretability of the translated output, is used for the quantitative evaluation on the task of semantic labels → cityscape photo. The intuition is that the FCN-8s, which is an off-the-shelf image segmentation approach, is more likely to accurately detect the semantic label if the translated image that fed to the FCN-8s is more realistic. Standard metrics for semantic segmentation are adopted.

Quantitative results that shown in Table 1 are consistent with those of the visual assessment in Figure 2. First, outperformance over the variation that without conditional adversarial loss is easy to be observed, since what the unconditional adversarial loss exactly penalizes is the unreality rather than the mismatching between the input and the output. When feature pyramids are utilized, different degree refinements are shown.

Table 1: FCN scores for different configurations with respect to the conditioning and feature pyramid, evaluated on the Cityscapes labels→photo translation task.

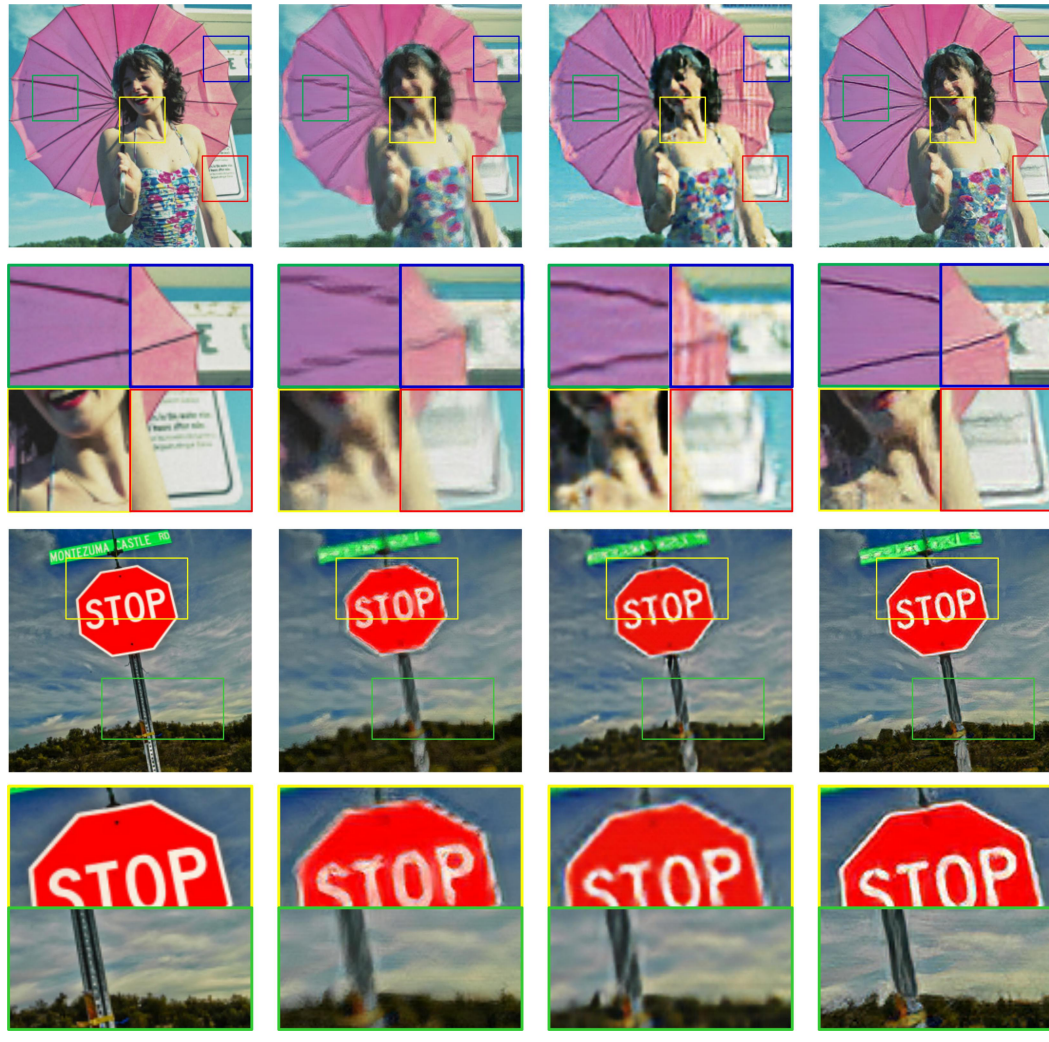
Loss	Per-pixel acc.	Per-class acc.	Class IOU
without conditional adversarial loss	0.42	0.10	0.07
without adversarial loss on low-frequency feature maps	0.71	0.25	0.18
without adversarial loss on high-frequency feature maps	0.45	0.15	0.11
completed model	**0.74**	**0.26**	**0.19**

We compare the performance of different variants with respect to the integration of the feature pyramids into the GAN framework. Evaluations have been conducted on the single image super-resolution task after being trained on the public dataset mscoco. [19] Widespread quantitative metrics such as Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are calculated since ground-true is available. Table 2 clearly reveals the outperformance of the completed model with all the metrics. To be more intuitive, visual performance are shown in Figure 3 while more results are given in the supplementary. Using the low-frequency feature pyramid alone produces even blurrier results than the traditional bicubic interpolation. Even though the multi-scale discriminators supply gradients to the generator, there is no steering for the generator to produce details when high-frequency information is filtered out by the low-frequency feature pyramid. In contrast, when only the high-frequency pyramid is utilized, the results are shown to be obviously sharper than training with its counterpart alone. Nevertheless, it sometimes results in sharpening artifacts since high-frequencies are undesirably intensified at the edges. Results yielded by the full integration are impressive even at a high magnification, as shown in Figure 3. Complementation between these two feature pyramids effectively enables the reduction of artifacts as well as the increase of realistic textures.

Table 2: Quantitative comparison of the ablation study with respect to the integration of different feature pyramids on x 4 super-resolution.

	without high-frequency pyramid	without low-frequency pyramid	Completed model
PSNR	20.34 dB	19.82 dB	**23.85 dB**
SSIM	0.8279	0.8097	**0.8563**

Indeed, our results are sometimes outperformed by some other approaches, such as the state-of-the-art super-resolution approach DRRN, [20] as shown in the supplementary. However, what we strive for is a better solution for general image-to-image translation tasks than the previous works rather than a best model for specific tasks. To further assess the dual-track pyramids, we perform another ablation study on the image inpainting task in the next section, in the purpose of comparing the best feature pyramid scheme against some state-of-the-art methods.



(a) (b) (c) (d)

Figure 3: Examples of the ablation study with respect to the integration of different feature pyramids on the single image super-resolution task. From left to right: (a) reference HR image; (b) without high-frequency pyramid; (c) without low-frequency pyramid; and (d) completed model.

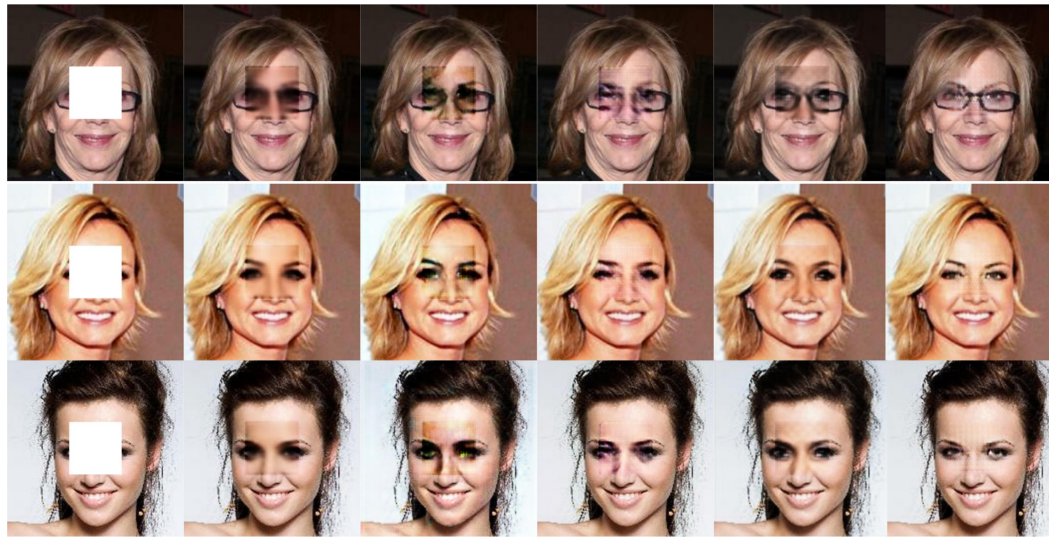
Comparisons against baselines

Recent state-of-the-art approaches that related to our methods are chose to be the baselines, including Pix2pix,[5] CE[16] and CNN. The CNN here is referred to the model that sets the perceptual feature loss and the two adversarial losses of our model to be zero; thus it is equivalent to a traditional CNN. For fair comparisons, this paper implements objectives of all the baselines with the same networks architecture and training details with us on each task unless specified otherwise.

The proposed approach is evaluated on the image inpainting task with holes too large to employ local non-semantic approaches. The purposes of this experiment are two. First, contrastive studies are conducted to learn the effect of our method. Second, we compare against several configurations of our approach that with different pyramids to choose the effective features. Experiment is conducted with data from the CelebA. More than 200k images are trained on the

aligned faces training set for 5 epochs. The training process is fast and converges in 10 hours, and the testing on 50k images only takes several minutes on a 1080 Ti GPU.

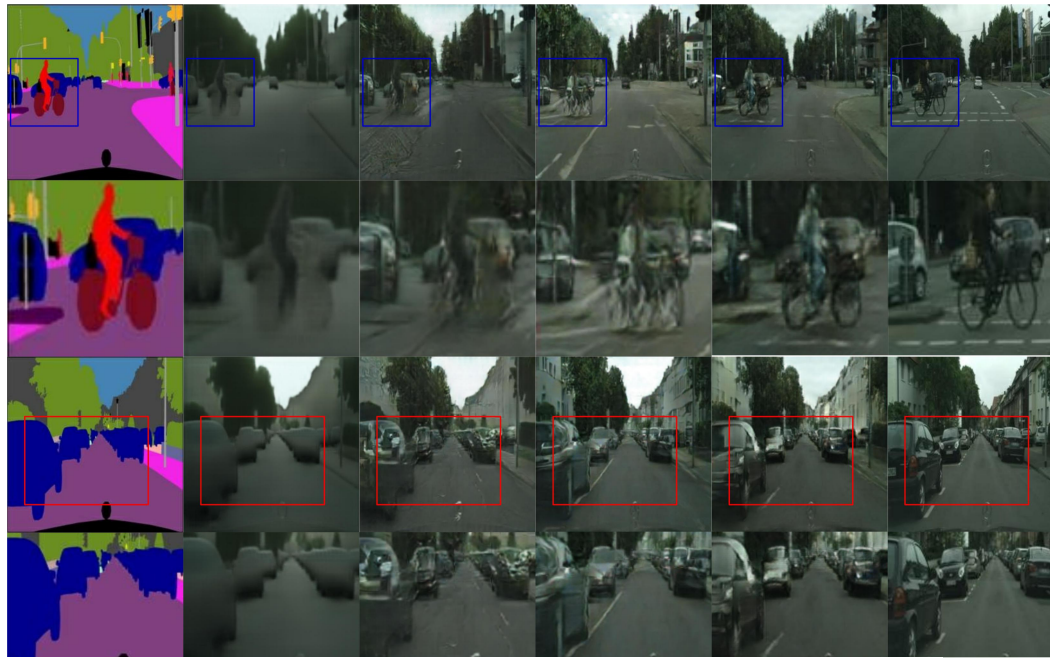
As shown in Figure 4, the results of CE [16] and pix2pix 5 are obviously inferior to ours as CE [16] predicts blurred central regions while the pix2pix [5] introduces unsatisfying artifacts. Then we proceed to validate the proposed coupled pyramid by comparisons against the alternative configurations, in particular, employing only a single low-frequency feature pyramid or its high-frequency counterpart. Figure 4 shows that the filled regions of (d) are lack of fine details. On the other hand, (e) tends to produce images with color distortion and checkerboard artifacts. Reasonable explanation might be that, even though the low-frequency feature matching is able to capture the overall appearance of the blank region, it tends to induce the averaging of potential location of details thus leads to over-smoothed images. On the other hand, heavy emphasis on high-frequency feature along with adversarial loss is responsible for obvious artifacts. Our completed model, which integrates the perceptual discrepancy into a cGAN model that with coupled feature pyramids, not only effectively sharpens the prediction but facilitates the hyper parameters tuning since adversarial network is unstable and sensitive to coefficient tuning in our preliminary experiment.



(a) (b) (c) (d) (e) (f)

Figure 4: Different models and several variants of our method for image inpainting. From left to right: (a) input; (b) Context-Encoder; [16] (c) pix2pix; [5] (d) ours model with a single low-frequency feature pyramid; (e) ours model with a single high-frequency feature pyramid; and (f) our completed model.

Then the proposed approach is evaluated on the tasks of Cityscapes labels→photo. Figure 5 shows several examples on the Cityscapes dataset.[18] While pix2pix [5] is able to produce much shaper results than those of CNN by means of introducing the conditional adversarial loss, artifacts can be easily observed. Even though CRN [14] produces less artifacts than pix2pix [5], in fact the transformed images of CRN [14] include hallucinated objects, such as the triple-wheel bicycle, as can been seen in the zoomed region. On the whole, none of the baselines above is competitive with our approach. Thus it is suggested that the perceptual adversarial loss and the separated feature pyramids are both essential to our superior performance. For better comparison, zoomed versions of certain regions-of-interest are shown below the outputs.



(a) (b) (c) (d) (e) (f)

Figure 5: Examples of different approaches to mapping Cityscapes labels to photo. From left to right: (a) input; (b) CNN; (c) pix2pix;[5] (d) CRN;[14] (e) ours; and (f) target.

CONCLUSION AND FUTURE WORK

This work develops a perceptual cGAN-based framework with feature pyramids for image-to-image translation. Feature maps with different frequency are separated to depress the emphasized on the sharpness, thus to alleviate the high-frequency artifacts. Benefits of traditional pixel-wise reconstruction loss and the perceptual loss are combined to capture both the appearance and spatial structure. Ablation experiments demonstrate that our loss function is superior to those without feature pyramids or perceptual loss and can become an effective tool for diverse image translation tasks. Compelling results of our model show that neither the features nor the priors need to be hand-engineered by our proposed method. Even though our results are outperformed by some task-specific approaches in some cases, we show great generality on a variety of tasks as well as the improvement on previous universal frameworks. Examples are shown in the supplementary.

Nevertheless, many issues still need to be further explored. When translating the cityscape photos into semantic labels using unpaired training data, the proposed approach sometimes rearranges the labels for trees and buildings. Ambiguity might be alleviated by leveraging a tiny fraction of annotated data or weakly supervised labeling for better prior. Second, while the proposed algorithm often succeeds in one-to-one mappings, extensive applications require multimodal outputs and variation technics seems to be a feasible solution to produce diverse results. Third, artifacts tend to be obvious when building up the images from lower resolution to a higher one, which is one of the typical issues of GANs. Since the deconvolution operation usually induces uneven overlaps on two axes of the images, creating checkerboard-like artifacts is hard to completely avoid. An alternative way to upsample is desired to resist the high-frequency features from being intensified. In addition, a better architecture of the perceptual feature extractor, such as the introspective neural networks or the cascading convolutional neural networks, to further improve the feature learning will be part of the future work.

REFERENCES

1. J. Johnson, A. Alahi, and F. F. Li, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” pp. 694–711, 2016.
2. A. Dosovitskiy and T. Brox, “Generating images with perceptual similarity metrics based on deep networks,” in Advances in Neural Information Processing Systems, 2016, pp. 658–666.
3. I. Goodfellow et al., “Generative adversarial nets,” in Advances in neural information processing systems, 2014, pp. 2672–2680.
4. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” arXiv preprint arXiv:1703.10593, 2017.
5. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv:1611.07004, 2016.
6. M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.
7. Y. Taigman, A. Polyak, and L. Wolf, “Unsupervised cross-domain image generation,” arXiv preprint arXiv:1611.02200, 2016.
8. L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.
9. C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” arXiv preprint arXiv:1609.04802, 2016.
10. Z. Yi, Hao (Richard) Zhang, P. Tan, and M. Gong, “DualGAN: Unsupervised Dual Learning for Image-to-Image Translation,” In ICCV, 2017 Oct 1, pp. 2868-2876.
11. T. Kim, M. Cha, H. Kim, J.K. Lee and J. Kim, “Learning to discover cross-domain relations with generative adversarial networks,” arXiv preprint arXiv:1703.05192.
12. M.Y. Liu and O. Tuzel. “Coupled generative adversarial networks,” In Advances in neural information processing systems, 2016, pp. 469-477.
13. T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs,” arXiv preprint arXiv:1711.11585, 2017.
14. Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” arXiv preprint arXiv:1707.09405, 2017.
15. P. J. Burt and E. H. Adelson, “The Laplacian Pyramid as a Compact Image Code,” IEEE Transactions on Communications, vol. 31, no. 4, pp. 532–540, 2003.
16. D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
17. A. Mahendran and A. Vedaldi, “Understanding deep image representations by inverting them,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5188–5196.
18. M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
19. T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, “Microsoft coco: Common objects in context,” In European conference on computer vision 2014 September, pp. 740-755.
20. Y. Tai, J. Yang, and X. Liu. "Image super-resolution via deep recursive residual network." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 2, p. 5. 2017.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61602413), the Natural Science Foundation of Zhejiang Province, China (No. LY19F030016) and the EU H2020 project-AniAge (No.691215).

ABOUT THE AUTHORS

Zhuorong Li received a PhD in Control Science and Engineering from Zhejiang University of Technology. She is currently a lecturer at the Zhejiang University City College. Her research interests include computational vision, deep learning and corresponding applications, machine learning. Contact her at lizr@zucc.edu.cn.

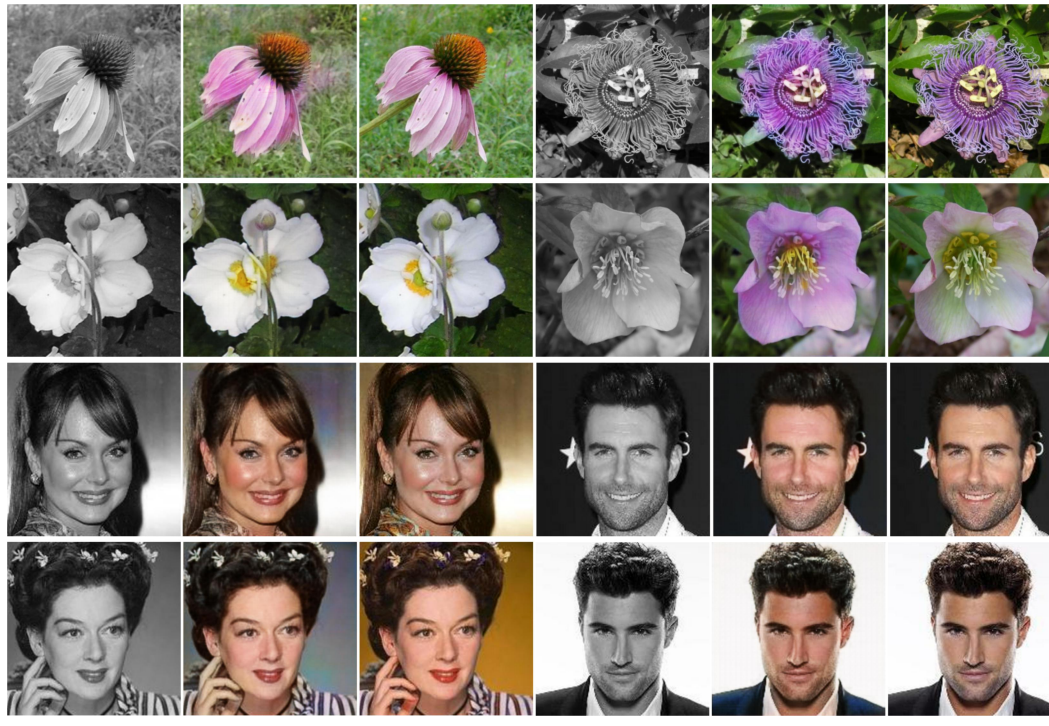
Minghui Wu received a PhD in Computer Science and Technology from Zhejiang University. He is currently a professor and the Dean of Computer and Computing Science School at the Zhejiang University City College. His research interests include mobile computing and artificial intelligence. Contact him at mhwu@zucc.edu.cn.

Jianwei Zheng received a PhD from Zhejiang University of Technology. He is currently an associate professor at the Zhejiang University of Technology. His research interests include machine learning and artificial intelligence. Contact him at zjw@zjut.edu.cn.

Hongchuan Yu is a Principal Academic of computer graphics in National Centre for Computer Animation, Bournemouth University. His specialties include geometry, graphics and image processing. He received his PhD in Computer Vision, Inst. of Intelligent Machine, Chinese Academy of Sciences, in 2000. He is a Member of IEEE and a fellow of High Education of Academy United Kingdom. Contact him at hyu@bournemouth.ac.uk.

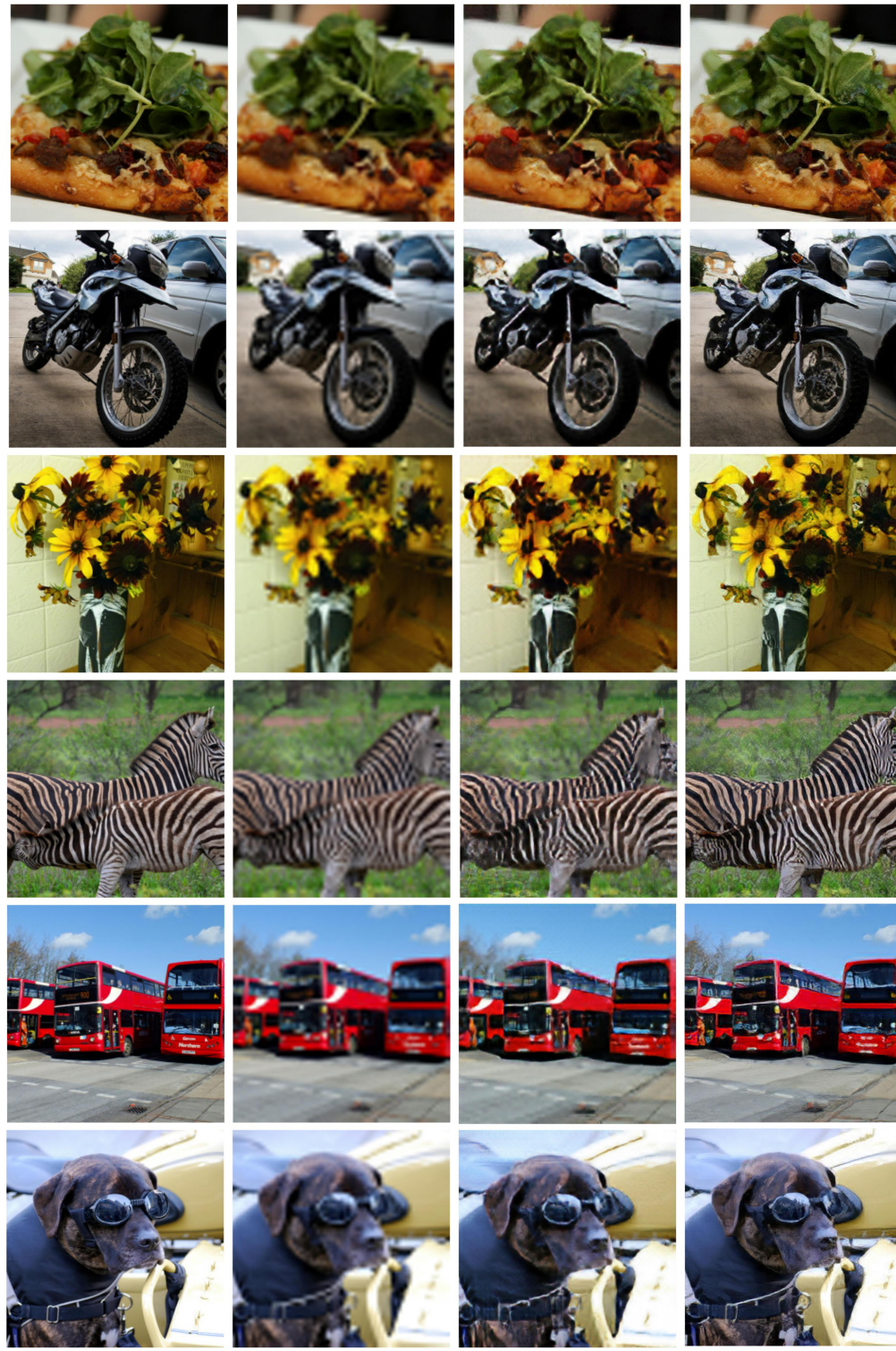
SUPPLEMENTARY

(The following images are best viewed and compared zoomed in.)



(a) (b) (c) (d) (e) (f)

Figure 6: Examples of image colorization by our method. Input ((a) and (d)); our results ((b) and (e)); and the reference images ((c) and (f)).



(a) (b) (c) (d)

Figure 7: Results of 4X super-resolution. From left to right: reference HR image, reconstruction results with corresponding to Bicubic, our method and the state-of-the-art approach DRRN. [20]

(a) (b) (c) (d)

Figure 8: Examples of our method for image inpainting. From left to right: (a) input; (b) Context-Encoder; [16] (c) pix2pix; [5] and (d) ours.

(a) (b) (c) (d) (e) (f)

Figure 9: Examples of image de-raining or de-snowing by our method. Input ((a) and (d)); our results ((b) and (e)); and the reference images ((c) and (f)).