WAVELET-BASED UNSUPERVISED LABEL-TO-IMAGE TRANSLATION

George Eskandar, Mohamed Abdelsamad, Karim Armanious, Shuai Zhang, Bin Yang

University of Stuttgart, Institute of Signal Processing and System Theory, Stuttgart, Germany

ABSTRACT

Semantic Image Synthesis (SIS) is a subclass of image-to-image translation where a semantic layout is used to generate a photorealistic image. State-of-the-art conditional Generative Adversarial Networks (GANs) need a huge amount of paired data to accomplish this task while generic unpaired image-to-image translation frameworks underperform in comparison, because they color-code semantic layouts and learn correspondences in appearance instead of semantic content. Starting from the assumption that a high quality generated image should be segmented back to its semantic layout, we propose a new Unsupervised paradigm for SIS (USIS) that makes use of a self-supervised segmentation loss and whole image wavelet based discrimination. Furthermore, in order to match the high-frequency distribution of real images, a novel generator architecture in the wavelet domain is proposed. We test our methodology on 3 challenging datasets and demonstrate its ability to bridge the performance gap between paired and unpaired models.

Index Terms— Semantic Image Synthesis, Unsupervised Training, GANs, Wavelet Transform

1. INTRODUCTION

Semantic image synthesis (SIS) is the task of generating high resolution images from user-specified semantic layouts. SIS opens the door to an extensive range of applications such as content creation and semantic manipulation by editing, adding, removing or changing the appearance of an object. By allowing concept artists and art directors to brainstorm their designs efficiently, it can play a pivotal role in graphics design. In addition, SIS can be used as a data augmentation tool for deep learning models, by generating training data conditioned on desired scenarios which might be hard to capture in real-life (e.g. corner cases in autonomous driving like accidents). One of the first SIS frameworks [2] was an adaptation of GANs [3–8]. SIS with SPatially Adaptive DEnormalization or SPADE [9] was proposed as an enhanced supervised methodology tailored specifically to suit SIS. Since then, a series of progressively enhanced frameworks were introduced [2, 9–12]. Most notably, OASIS [13] extends the SPADE framework with a novel discriminator design and achieves state-of-the-art in supervised SIS.

However, the problem of semantic image synthesis has mostly been addressed in a supervised setup. Although state-of-the-art methods [2, 9–13] can produce visually appealing high resolution images, they still depend on the availability of annotated training data which is expensive to acquire. For instance, the average annotation time for one image in the Cityscapes dataset is 1 hour [14].

Unpaired conditional GAN frameworks [15–25] can be used to bypass the need for paired training data in SIS, but they suffer from several drawbacks: (1) instead of using one-hot encoding, these models color-code each class in the input semantic layout, which creates an artificial mapping between layouts and images, (2) the unsupervised losses in these works force relationships between the labels and images that do not preserve the semantic content in the case of SIS and (3) the normalization layers in the architecture of unsupervised models wash away the semantic labels as noted in [9]. Consequently, the generated samples from these models suffer from poor quality, especially when the number of classes in the training dataset is too big.

In this work, we propose an unsupervised framework which can synthesize realistic images from labels without the use of paired data. This is a first step towards bridging the performance gap between paired and unpaired frameworks. By virtue of its design, the unpaired setting can help eliminate dataset biases and push the model towards a better multimodal generation. The proposed Unsupervised SIS (USIS) [1] is a paradigm which involves an adversarial training between a generator and a whole image wavelet-based discriminator, and a cooperative training between the generator and a UNet segmentation network [26]. More precisely, the discriminator fosters the generator to match the distribution of the real images while the UNet gives a pixel-level feedback to the generator by means of a cross entropy loss. In addition, upon observing that convolutional networks are biased towards low-frequencies [27–37], we provide the wavelet decomposition of the real and fake images as input to our discriminator and we propose a novel wavelet-based generator architecture in order to approximate the high-frequency distribution of real images. We perform extensive experiments on 3 image datasets (COCO-stuff [38], Cityscapes [14] and ADE20K [39]) in an unpaired setting to showcase the ability of our model to generate a high diversity of photorealistic images.

An extended version of this paper is available in `https://arxiv.org/abs/2109.14715` [1].

2. METHOD

In SIS, we seek to synthesize an RGB image $\mathbf{x}$ from a semantic mask $\mathbf{m}$ (one-hot encoded) with C classes in an unsupervised way. To achieve this, we propose the USIS framework which consists of three parts: (1) a waveletSPADE Generator $\mathcal{G}$, (2) a wavelet Discriminator $\mathcal{D}$ and (3) a UNet segmentation network $\mathcal{S}$. The generator generates a synthesized image $\hat{x}$ from the semantic map $\mathbf{m}$, the discriminator makes a decision whether the generated image is real or fake while the segmentation network tries to segment the generated image back to the mask $\mathbf{m}$. $\mathcal{S}$ only observes the generated images unlike the discriminator which sees real and generated images. In the following, we are going to introduce each of the 3 components in detail. The paradigm is depicted in Figure 1.

2.1. Semantic consistency loss

A photorealistic image contains different objects which have different appearances and semantic meanings. Starting from this assumption, we impose a semantic consistency constraint to force the generator to produce images that lie on the real image manifold. A class-balanced segmentation loss $\mathcal{L}_{seg}$ between the synthesized images and the input labels is proposed to punish the inseparability between regions belonging to different semantic labels. This loss is self-supervised because it uses the input semantic layout as the groundtruth for segmentation, similar to autoencoders. The self-supervised loss pushes the generator to synthesize small classes and achieve a better semantic alignment. It can be expressed as:

$$\mathcal{L}_{seg} = -\mathbb{E}_{\mathbf{m}}\left[\sum_{c=1}^{C}\alpha_c \sum_{i,j}^{H\times W} \mathbf{m}_{c,i,j}\log(\mathcal{S}(\mathcal{G}(\mathbf{m}))_{c,i,j})\right] \quad (1)$$

The class-balancing weights α_c are proportional to the inverse of the per-pixel class-frequency.

$$\alpha_c = \frac{H\times W}{\sum_{i,j}^{H\times W}\mathbb{E}_{\mathbf{m}}\left[\mathbb{1}[\mathbf{m}_{c,i,j}]\right]} \quad (2)$$

The overall loss can be expressed as $\mathcal{L}_{seg} + \lambda\mathcal{L}_{adv}$, where λ is a hyperparameter.

2.2. Whole image wavelet-based discrimination

The discriminator is an essential part of the framework because it is responsible for capturing the data statistics. Most importantly, it prevents the generator from learning trivial mappings (like identity mapping) that minimize the self-supervised segmentation loss. We propose to use whole image wavelet based discrimination for this task. More specifically, we incorporate the SWAGAN discriminator architecture, which was previously proposed in [8] to enhance the texture of generated images. By allowing the discriminator to process the Discrete Wavelet Transform (DWT) of the image, the higher frequencies are not entirely lost in the downsampling layers of the discriminator. Moreover, small objects in the spatial domain have wavelet coefficients with bigger magnitude, which are now observed by the discriminator and have a bigger contribution in $\mathcal{L}_{adv}$. Formally, we define channelwise-DWT as the operation that takes a tensor of dimensions $(c \times w \times h)$ and outputs wavelet coefficients of dimensions $(4c \times w/2 \times h/2)$, where each input channel is transformed into 4 subbands of lower resolution, which are concatenated channelwise. Spatial-DWT refers to the concatenation of the subbands along the height and width of the features, yielding a tensor of the same dimension as the input. The Inverse Wavelet Transform (IWT) is defined as the inverse operation, which maps form the wavelet domain to the spatial domain. Unless otherwise stated, DWT and IWT refer to the channelwise-DWT and channelwise-IWT respectively. The 2 arrangements can be seen in Figure 2. More details about the architecture of the utilized discriminator can be found in our prior publication [1].

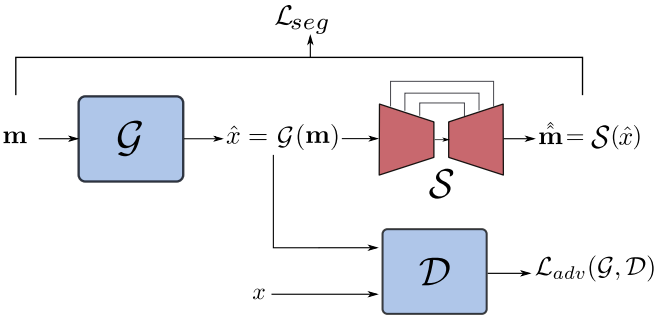
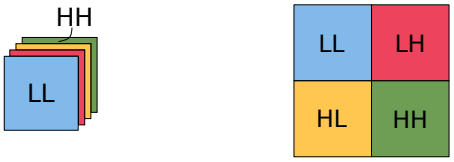


Fig. 1: The proposed unsupervised SIS paradigm



(a) Channelwise-DWT (b) Spatial-DWT

Fig. 2: Different arrangement of the wavelet subbands inside the network

2.3. Wavelet-based generator design

In this subsection, we present a novel generator architecture: the waveletSPADE which is built upon the SPADE generator but designed to generate wavelet coefficients directly instead of pixels. More concretely, two modules are introduced: the wavelet upsample (WU) and the pixelSPADE (PS). The waveletSPADE architecture inherits its decoder-structure from SPADE but learns all the features in the wavelet domain. The input is the one-hot encoded segmentation map $\mathbf{m}$ concatenated with a 3D noise tensor, $\mathbf{z}$. Introduced in OASIS [13, 40], 3D noise is a technique to learn multimodal generation by injecting structured noise in multiple layers of the network. The tensor is formed by sampling a latent vector

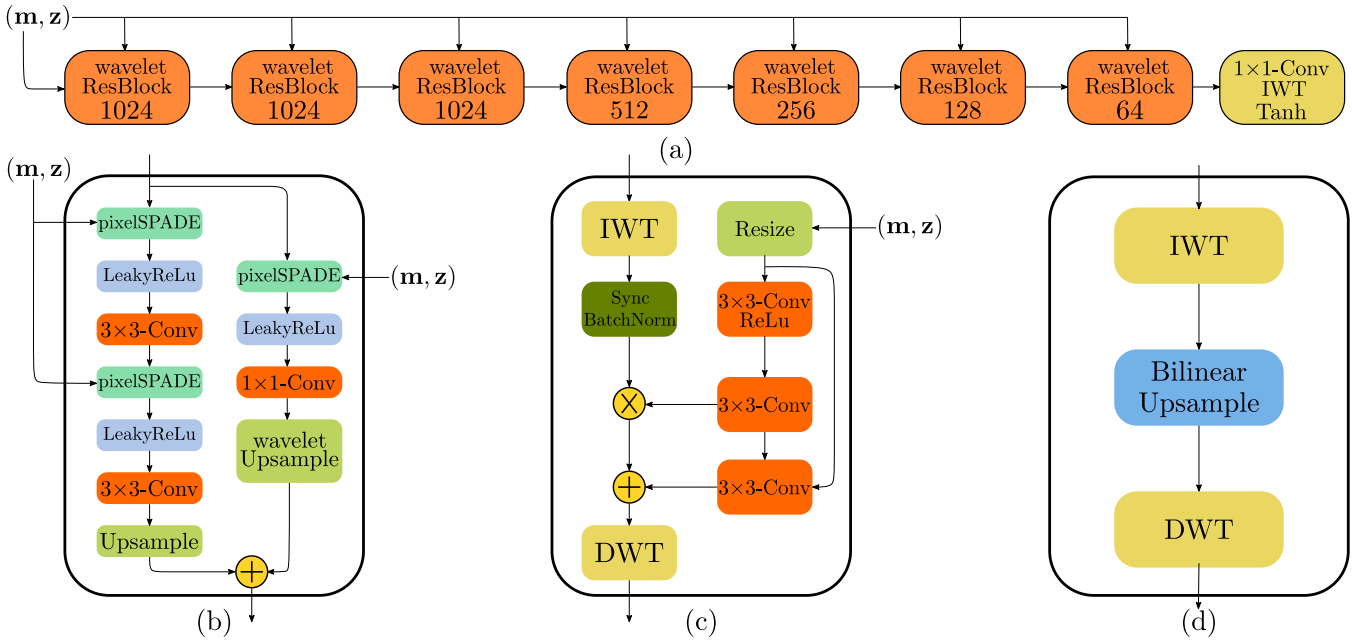


Fig. 3: The proposed waveletSPADE generator architecture (a) builds upon the OASIS generator [13] by replacing the ResBlocks with the depicted waveletResBlock (b) which encompasses the novel pixelSPADE (c) as well as the WU blocks (d). The number of channels per layer is shown in each waveletResBlock.

of dimension Z from a zero-mean unit-variance Gaussian distribution and propagating it at each pixel of the segmentation map resulting in a shape of $((C + Z) \times H \times W)$. The architecture and the blocks are shown in Figure 3.

Wavelet Upsampling: In contrast to the original SPADE architecture, ResBlocks and upsampling layers are not interleaved. Instead, we move the upsampling layers inside the ResBlocks but use different scaling methods inside each branch: the identity branch is upsampled using a waveletUpsample (WU) block, while the residual branch is upsampled with a nearest-neighbour interpolation. The motivation for such a design choice is that we would like to preserve identity mappings in the shortcut connection of the ResBlock, as they are necessary to channel proper gradients to the early layers of the network which produce coarse features. The waveletUpsample layer, inspired by SWAGAN[8], consists of an IWT, followed by a bilinear upsampling in the spatial domain and a DWT. This stands in contrast to applying transposed convolutions directly on high features inside the network because this would ignore that the 4 subbands of DWT contain different frequency information and would introduce low-frequency artifacts in the generated image. On the other hand, we use nearest-neighbour upsampling in the residual branch because it preserves the sharpness of the wavelet coefficients (otherwise high value coefficients could be attenuated by linear interpolation techniques).

The **pixelSPADE** layer replaces the original SPADE layer in all the ResBlocks. Since the style of an image correlates to the mean and standard deviation of the features inside the network, the original SPADE was designed to generate demodulation parameters per pixel. However, applying SPADE to each subband of the DWT is suboptimal as SPADE was designed to be applied in the spatial domain, not on the frequency components of the image. Therefore, we propose to use pixelSPADE which consists of an IWT operation, a SPADE layer and a DWT. The block thus offers the advantage of preserving the frequency content while applying the style to the features in the spatial domain.

3. EXPERIMENTS

We conduct our experiments for unsupervised SIS on 3 datasets: Cityscapes[14], COCO-stuff[38] and ADE20K[39]. Cityscapes contains street scenes in German cities with pixel-level annotations of 19 classes. It is widely used for vision tasks in autonomous driving and contains 3000 training images and 500 test images. ADE20K and COCO-stuff are more challenging datasets because they offer a high diversity of indoor and outdoor scenes; and they have a lot of semantic classes. COCO-stuff has 182 classes while ADE20K has 150 classes. These 3 datasets are the standard benchmark in supervised SIS. We use a resolution of 256 $\times$ 512 for Cityscapes and a resolution of 256 $\times$ 256 for ADE20K and Cocostuff. In previous works on unpaired GANs, the semantic image synthesis experiments were only performed on Cityscapes and/or datasets with small number of classes. The standard evaluation metrics for this task are utilized to measure both the quality and diversity of generated images. Specifically, we use the Frechet Inception Distance or FID[41], to assess both quality and diversity. We also follow SPADE [9] and run

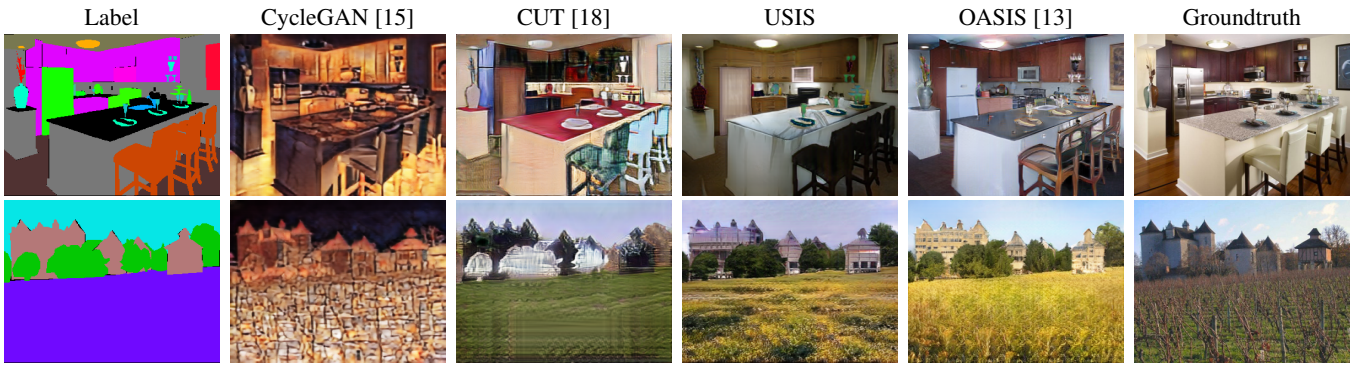


Fig. 4: Qualitative comparison against state-of-the-art unpaired SIS on ADE20K. OASIS is the state-of-the-art in paired SIS.

Method	Cityscapes		ADE20K		COCO-stuff	
	FID↓	mIoU↑	FID↓	mIoU↑	FID↓	mIoU↑
CycleGAN [15]	87.2	24.5	96.3	5.4	104.7	2.08
MUNIT [16]	84	8.2	n/a	n/a	n/a	n/a
DRIT [17]	164	9.5	132.2	0.016	135.5	0.008
Distance [20]	78	17.6	80	0.035	92.4	0.014
GCGAN [19]	80	8.4	92	0.07	99.8	0.019
CUT [18]	57.3	29.8	79.1	6.9	85.6	2.21
USIS	**50.14**	**42.32**	**34.5**	**16.95**	**28.6**	**13.4**

Table 1: Comparison against state-of-the-art unpaired GANs.

Method	FID	mIoU
OASIS	52.19	**42.8**
OASIS + IWT	50.52	40.27
OASIS + IWT + WU	51.37	39.88
OASIS + IWT + WU*	54.78	35.74
OASIS + Spatial-IWT + WU	55.44	28.22
OASIS + Spatial-IWT + WU + PS	52.69	40.66
OASIS + IWT + WU + PS	**50.14**	42.32

Table 2: Ablation study on Cityscapes.

pretrained semantic segmentation models [42–44] on the generated images and report the mean Intersection-over-Union (mIoU) to evaluate the semantic alignment and visual quality. We also perform an ablation study on Cityscapes to analyze the effect of the different components in the generator architecture. All the experiments are conducted using the OASIS generator (SPADE generator + 3D noise), the wavelet-based discriminator and the UNet. The batchsize is 8. Adding an IWT after the generator is referred to as OASIS + IWT. WU* refers to the proposed wavelet upsample in [45].

4. RESULTS

The quantitative comparison is reported in Table 1 and some samples of the generated images are shown in Figure 4. CycleGAN results in the most unrealistic images, as it is unable to math the color and texture distribution of the real images although it is able to produce objects with clear boundaries. This is reflected in the suboptimal FID scores but relatively high mIoU scores in all 3 datasets in Table 1. More recent frameworks like DistanceGAN or GCGAN have low FID scores but the objects might often be indistinguishable. CUT has the best FID and mIoU combination among the baselines, but suffers from a deterioration when the number of classes is high. The proposed USIS model is able to bypass this problem, and generates photorealistic images on the challenging datasets by virtue of the semantic consistency loss and its wavelet-based discriminator and generator designs.

The results of the ablation study are reported in Table 2. The simple addition of a channelwise IWT at the end of the OASIS generator enhances the FID as the generator is better able to generate more high-frequency content and produce better texture. However, this comes at the expense of the semantic alignment. Replacing upsampling layers by WU layers alone seems to slightly worsen the results, leading to more misalignment. It's the introduction of the PixelSPADE layer that restores the good alignment results with a negligible degradation while lowering the FID, thus approaching the state-of-the-art supervised performance (47.7 in OASIS [13]). We have also experimented with Spatial-DWT and IWT leading to suboptimal results. In contrast, representing frequency information in the channels helps the network learn more useful features.

5. CONCLUSION

In this work, a framework for semantic image synthesis in an unpaired setting was proposed (USIS). It deploys a waveletSPADE generator along with a UNet and an unconditional whole image wavelet-based discriminator. The UNet fosters class separability and content preservation while the discriminator matches the color and texture distribution of real images. The effectiveness of the proposed framework in the semantic image synthesis was shown on 3 challenging datasets: Cityscapes, ADE20K and Cocostuff. USIS outperformed prior unpaired GANs while approaching the performance of supervised frameworks. An ablation study was performed to analyze the role of the different components in the unsupervised paradigm.

Acknowledgement

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project "AI Delta Learning". The authors would like to thank the consortium for the successful cooperation.

References

[1] George Eskandar et al. *USIS: Unsupervised Semantic Image Synthesis*. 2021. arXiv: 2109.14715 [cs.CV].

[2] Ting-Chun Wang et al. "High-resolution image synthesis and semantic manipulation with conditional GANs". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[3] I. Goodfellow et al. "Generative Adversarial Nets". In: *NIPS*. 2014.

[4] Tero Karras et al. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*. http://arxiv.org/abs/1710.10196. 2018.

[5] Tero Karras, S. Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4396–4405.

[6] Tero Karras et al. "Analyzing and Improving the Image Quality of StyleGAN". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 8107–8116.

[7] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[8] Rinon Gal et al. "SWAGAN: A Style-based Wavelet-driven Generative Model". In: *ArXiv* abs/2102.06108 (2021).

[9] Taesung Park et al. "Semantic image synthesis with spatially-adaptive normalization". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[10] Zhentao Tan et al. "Rethinking Spatially-Adaptive Normalization". In: *arXiv:2004.02867* (2020).

[11] Xihui Liu et al. "Learning to predict layout-to-image conditional convolutions for semantic image synthesis". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[12] Peihao Zhu et al. "SEAN: Image Synthesis With Semantic Region-Adaptive Normalization". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 5103–5112.

[13] Edgar Schönfeld et al. "You Only Need Adversarial Supervision for Semantic Image Synthesis". In: *International Conference on Learning Representations*. 2021.

[14] Marius Cordts et al. "The cityscapes dataset for semantic urban scene understanding". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[15] Jun-Yan Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *International Conference on Computer Vision (ICCV)*. 2017.

[16] Xun Huang et al. "Multimodal unsupervised image-to-image translation". In: *European Conference on Computer Vision (ECCV)*. 2018.

[17] Hsin-Ying Lee et al. "Diverse Image-to-Image Translation via Disentangled Representation". In: *European Conference on Computer Vision (ECCV)*. 2018.

[18] Taesung Park et al. "Contrastive Learning for Unpaired Image-to-Image Translation". In: *European Conference on Computer Vision*. 2020.

[19] Huan Fu et al. "Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[20] Sagie Benaim and Lior Wolf. "One-sided unsupervised domain mapping". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[21] Yaniv Taigman, Adam Polyak, and Lior Wolf. "Unsupervised Cross-Domain Image Generation". In: *International Conference on Learning Representations (ICLR)*. 2017.

[22] Ashish Shrivastava et al. "Learning from simulated and unsupervised images through adversarial training". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[23] Konstantinos Bousmalis et al. "Unsupervised pixel-level domain adaptation with generative adversarial networks". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[24] Matthew Amodio and Smita Krishnaswamy. "Travelgan: Image-to-image translation by transformation vector learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 8983–8992.

[25] Rui Zhang, Tomas Pfister, and Jia Li. "Harmonic unpaired image-to-image translation". In: *International Conference on Learning Representations (ICLR)*. 2019.

[26] Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. "A U-Net Based Discriminator for Generative Adversarial Networks". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[27] Yuanqi Chen et al. "SSD-GAN: Measuring the Realness in the Spatial and Spectral Domains". In: *AAAI*. 2021.

[28] Ricard Durall, Margret Keuper, and Janis Keuper. "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 7887–7896.

[29] Tarik Dzanic and F. Witherden. "Fourier Spectrum Discrepancies in Deep Network Generated Images". In: *ArXiv* abs/1911.06465 (2020).

[30] Xing Gao and H. Xiong. "A hybrid wavelet convolution network with sparse-coding for image super-resolution". In: *2016 IEEE International Conference on Image Processing (ICIP)* (2016), pp. 1439–1443.

[31] P. Liu et al. "Multi-Level Wavelet Convolutional Neural Networks". In: *IEEE Access* 7 (2019), pp. 74973–74985.

[32] Travis Williams and Robert Y. Li. "Wavelet Pooling for Convolutional Neural Networks". In: *ICLR*. 2018.

[33] Lin Liu et al. "Wavelet-Based Dual-Branch Network for Image Demoireing". In: *ArXiv* abs/2007.07173 (2020).

[34] Eunhee Kang, Junhong Min, and J. C. Ye. "A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction". In: *Medical Physics* 44 (2017), e360–e375.

[35] Yunfan Liu, Qi Li, and Zhenan Sun. "Attribute-Aware Face Aging With Wavelet-Based Generative Adversarial Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 11869–11878.

[36] Jianyi Wang et al. "Multi-level Wavelet-based Generative Adversarial Network for Perceptual Quality Enhancement of Compressed Video". In: *ArXiv* abs/2008.00499 (2020).

[37] Huaibo Huang et al. "Wavelet Domain Generative Adversarial Network for Multi-scale Face Hallucination". In: *International Journal of Computer Vision* 127 (2019), pp. 763–784.

[38] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. "Cocostuff: Thing and stuff classes in context". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[39] Bolei Zhou et al. "Scene parsing through ade20k dataset". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[40] Yazeed Alharbi and Peter Wonka. "Disentangled Image Generation Through Structured Noise Injection". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[41] Martin Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[42] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. "Dilated residual networks". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[43] Liang-Chieh Chen et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs". In: *International Conference on Learning Representations (ICLR)* (2015).

[44] Tete Xiao et al. "Unified perceptual parsing for scene understanding". In: *European Conference on Computer Vision (ECCV)*. 2018.

[45] P. Tsai and T. Acharya. "Image Up-Sampling Using Discrete Wavelet Transform". In: *JCIS*. 2006.