Deep Residual Transform for Multi-scale Image Decomposition

Yuhao Chen University of Waterloo
Alexander Wong University of Waterloo
Yuan Fang University of Waterloo
Yifan Wu University of Waterloo
Linlin Xu University of Waterloo
Email: {yuhao.chen1, alexander.wong, yuan.fang, yifan.wu1, l44xu}@uwaterloo.ca

Abstract

Multi-scale image decomposition (MID) is a fundamental task in computer vision and image processing that involves the transformation of an image into a hierarchical representation comprising of different levels of visual granularity from coarse structures to fine details. A well-engineered MID disentangles the image signal into meaningful components which can be used in a variety of applications such as image denoising, image compression, and object classification. Traditional MID approaches such as wavelet transforms tackle the problem through carefully designed basis functions under rigid decomposition structure assumptions. However, as the information distribution varies from one type of image content to another, rigid decomposition assumptions lead to inefficiently representation, i.e., some scales can contain little to no information. To address this issue, we present Deep Residual Transform (DRT), a data-driven MID strategy where the input signal is transformed into a hierarchy of non-linear representations at different scales, with each representation being independently learned as the representational residual of previous scales at a user-controlled detail level. As such, the proposed DRT progressively disentangles scale information from the original signal by sequentially learning residual representations. The decomposition flexibility of this approach allows for highly tailored representations cater to specific types of image content, and results in greater representational efficiency and compactness. In this study, we realize the proposed transform by leveraging a hierarchy of sequentially trained autoencoders. To explore the efficacy of the proposed DRT, we leverage two datasets comprising of very different types of image content: 1) CelebFaces and 2) Cityscapes. Experimental results show that the proposed DRT achieved highly efficient information decomposition on both datasets amid their very different visual granularity characteristics.

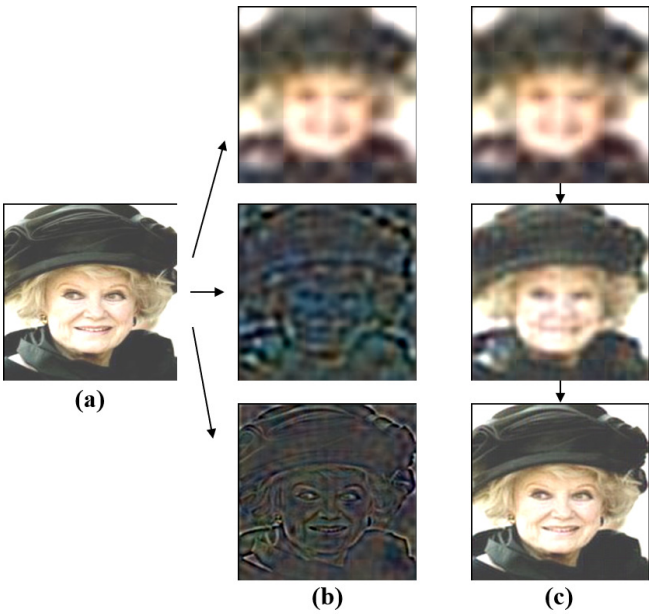

Fig. 1: Deep Residual Transform decomposes the input image (column a) into 3 bases (column b). Each basis characterizes the input signal at a particular scale. Column c shows the reconstructed image at each scale.

1 Introduction

Multi-scale image decomposition (MID) is a fundamental task in computer vision and image processing that involves the transformation of an image into a hierarchical representation comprising of different levels of visual granularity from coarse structures to fine details. A well-engineered MID disentangles the image signal into meaningful components which can be used in a variety of applications such as image denoising, image compression, and object classification. For example, in image denoising [1], signal are decomposed into scales at different frequency bands, and high frequency scales that often represents noise are thresholded or removed to improve visual quality. In the same way, image compression [2] removes high frequency scale to save storage bits while preserving information perceptually. In object classification and detection [3], MID provides additional hierarchical relationship along with the disentangled object features, making classification and detection easier.

Most traditional image decomposition methods built the representation on linear basis and non-data-driven basis [4–6], such as Discrete Cosine Transform (DCT), Wavelet Transform, and Pyramid Representations. DCT [4] encode images into frequency representation using summation of cosine functions. Pyramid representation [6] decomposes the image by iteratively applying a linear smooth kernel and subsampling the image. The smoothed images from all iterations create a multi-scale representation. Similarly, instead of using a single linear filter, Wavelet Transform [5] applies a family of invertible and orthogonal filters. In Wavelet Transforms, image is convolved with a high pass filter and a low pass filter at each scale. The scale information is captured by the high pass filter, and the response of the low pass filter is passed to the next scale. Among the decomposition methods, DCT lacks temporal information to represent non-stationary features such as edges, since its basis function has an infinite length. Pyramid based transforms and Wavelet Transforms captures both temporal and frequency information but both approaches rely on rigid decomposition structure assumptions and well engineered basis to band pass certain frequency content at each scale. Instead of using linear filters, several other studies [7–9] propose non-linear edge-preserving filters to smooth textures while keeping structural information from edges, but these filters tend to over-smooth object surfaces and create cartoon-like representations. As the information distribution varies from one type of image content to another, rigid decomposition assumption leads to inefficiently representation. For example, if a dataset only contains information at extreme low or high frequency, some scales representing the middle frequency band can contain little to no information. In addition, nature images consist of complex and non-linear features, which poses a significant challenge for the above mentioned linear MIDs to efficiently represent.

To address this issue, we present Deep Residual Transform, a data-driven MID strategy where the input signal is transformed into a hierarchy of non-linear representations. Each representation is independently learned as the representational residual of previous scales. In addition, this process allows user to control the level of details at each representational scale. As such, the proposed DRT progressively disentangles scale information from the original signal by sequentially learning residual representations. The decomposition flexibility of this approach allows for highly tailored representations cater to specific types of image content, and results in greater representational efficiency and compactness. In this study, we realize the proposed transform by leveraging a hierarchy of sequentially trained autoencoders. To explore the efficacy of the proposed DRT, we leverage two datasets comprising of very different types of image content: 1) CelebFaces and 2) Cityscapes. Experimental results show that the proposed DRT achieved highly efficient information decomposition on both datasets amid their very different visual granularity characteristics, particularly when studied via frequency analysis. Figure 1 illustrates a 3 level decomposition produced by DRT, and the representation at each scale provides unique visual granularity characteristics. Furthermore, we demonstrate the representation flexibility of the proposed DRT by showing how parameter

adjustments can control the coarseness/fineness at each scale.

2 Deep Residual Transform

In this section, we describe the derivation of the proposed Deep Residual Transform (DRT). We first model the input signal X as the summation of n bases that characterize the signal information with increasingly fine details:

$$X = \sum_{i=1}^{n} b_i \tag{1}$$

where b_i is the basis characterizing signal information at ith scale. Basis at lower scale represents coarse signal structures, while basis at higher scale represents fine details (e.g. b_0 represents the coarsest signal structures and b_n represents the finest details in the signal). The signal information represented by each basis describes a unique signal characteristic at a particular scale that is not contained in any other basis. By adding the bases up to a particular scale j, we can reconstruct a "coarse" version of the original signal, and we denote it as the reconstructed signal $\hat{X}_j$:

$$\hat{X}_j = \sum_{i=1}^{j} b_i \tag{2}$$

When the scale j is at n, we obtain a perfect reconstruction of the original signal X:

$$\hat{X}_{j=n} = \sum_{i=1}^{n} b_i = X \tag{3}$$

The purpose of developing DRT is to decompose a signal into a hierarchical representation through flexible basis learning. A good representation basis has two key properties, information preservation and representation uniqueness. Traditional approaches [5, 6] rely on rigid decomposition structure and orthogonal basis to define data-invariant coarseness/fineness boundaries between the scales, and thus achieve perfect reconstruction and prevent representation redundancy. Since using learnable basis means the boundaries between the scales vary with datasets, we take a different approach by leveraging a deep cascade residual framework that finds basis sequentially.

Let the initial residual r_0 to be input X:

$$r_0 = X = \sum_{i=1}^{n} b_i \tag{4}$$

We rearrange the equation for r_0 and obtain the following expression:

$$r_0 = b_1 + \sum_{i=2}^{n} b_i = b_1 + r_1 \tag{5}$$

where

$$r_1 = \sum_{i>1} b_i = \sum_{i=1}^{n} b_i - b_1 = X - \hat{X}_1 \tag{6}$$

and signal r_1 represents the reconstruction residual between the input X and the reconstructed signal $\hat{X}_1$ at scale 1. Similarly, we can rewrite the equation for residual r_1 to obtain the residual r_2 at scale 2:

$$r_1 = b_2 + \sum_{i=3}^{n} b_i = b_2 + r_2 \tag{7}$$

where

$$r_2 = \sum_{i>2} b_i = X - \hat{X}_2 \tag{8}$$

Generalizing the equations above, ith scale residual r_i can be obtained by taking the difference of the input X and the ith reconstructed signal $\hat{X}_i$:

$$r_i = X - \hat{X}_i \tag{9}$$

It can also be obtained by the summation of the $i+1$th scale basis b_{i+1} and $i+1$th scale residual r_{i+1}:

$$r_i = b_{i+1} + r_{i+1} \tag{10}$$

This equation shows that in order to obtain the basis b_{i+1} at next scale, we only need to know the residual r_i at scale i. Furthermore, as the basis b_{i+1} only characterizes signal information at a particular scale, any information beyond the scale is contained in its residual r_{i+1} for further disentanglement. We set the last basis b_n equals to the $n-1$th scale residual r_{n-1}, so that the last residual $r_n = 0$, and we have a perfect reconstructed signal.

Base on the above observations, we can safely assume that each basis can be independently learned and has no impact on the learning of other bases. Thus, we model the basis at scale i as a function of the residual at previous scale $i-1$:

$$b_i = f_i(r_{i-1}) \tag{11}$$

where f_i is the basis learning function for scale i. Combining the above expression with Equation 10, we obtain an expression to learn all basis for $i < n$:

$$r_i = r_{i-1} - f_i(r_{i-1}) \tag{12}$$

Finally, given the Equation 12, we have a deep cascading residual framework for DRT where each basis at a scale i is learned as the representational residual of previous scale $i-1$ by leveraging a basis learning function. This framework decomposes the input signal into a set of learned bases $B = \{b_1, b_2, ..., b_n\}$ which is organized in a hierarchical structure that characterizes signal information from coarse scale to fine scale. Figure 2 illustrates the deep cascading residual framework for DRT.

2.1 Realization of Deep Residual Transform via Hierarchy of Autoencoders

Recent advance in Deep Learning Models has achieved huge success on reconstructing and generating inputs using a non-linear encoder/decoder style representation [10–13]. In this paper, we use the simplest encoder/decoder representation, autoencoder [13], as our basis function for realizing DRT. Through empirical studies, we observe that compared to typical autoencoder with multi-layer encoder and decoder (multi-layer autoencoder), autoencoder with single encoder and decoder layer (2-layer autoencoder) provides more control to the reconstruction coarseness and fineness. Figure 3 describes the architecture of our 2-layer autoencoder. In particular, the learning progression of a 2-layer autoencoder (as shown in 4 top row) describes a non-linear transition from coarse structures to fine details for the representation, whereas the learning progression of a multi-layer autoencoder (as shown in 4 bottom row) only shows the illumination change of the representation. Therefore, by leveraging early stopping in the learning progression, we can control the level of details in a representation. However, utilizing learning progression alone is not enough to disentangle multi-scale information. The learning process starts with a coarse representation at a particular scale and converges to a finer representation at another scale. It is not guaranteed that the starting scale and the converging scale represents the most coarse structures and the finest details. In fact, the starting scale and the converging scale are heavily influenced by the parameters of autoencoder, that are the convolutional filter size and stride size. In the following, we describe the impact of each autoencoder parameter, particularly the decoder parameters. To simplify the design, we mirror the parameter setting of decoder to encoder except that decoder uses a convolutional transpose layer and encoder uses a convolutional layer.

- **Stride size** denoted as θ_S controls the sampling rate of the representation. It is the key parameter that controls the converging scale of representation learning (i.e. the learned representation will not characterizing any signal information beyond the converging scale).
- **Hidden Layer Channel Number** controls the representation variation. We observe that excessive channel number leads to learning representations that are pure noise and have minimal impact on the reconstructed signal. Therefore, we keep the channel number the same as the input channel number.
- **Filter size** denoted as θ_f controls starting scale of representation learning. Large filter size means autoencoder starts the learning process with coarse representation, and smaller filter size means autoencoder starts the learning process with fine representation.
- **Representation degree** denoted as θ_d is derived from filter size divided by the stride size.

$$\theta_d = \frac{\theta_f}{\theta_S} \tag{13}$$

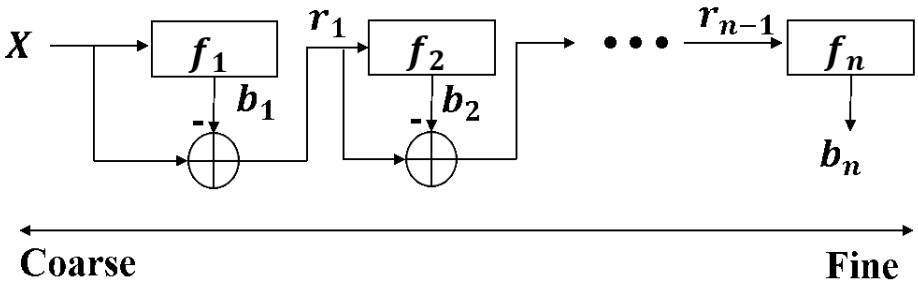

Fig. 2: Deep Residual Transform Framework

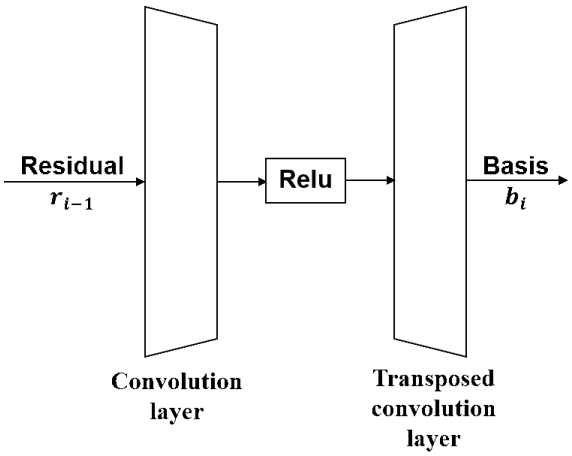

Fig. 3: Architecture of the 2-layer autoencoder for basis learning

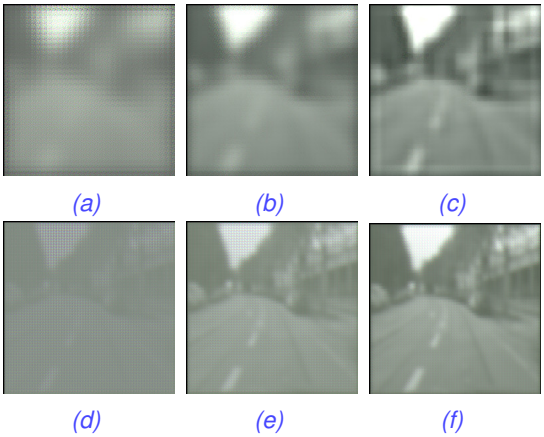

Fig. 4: Basis learning progression using autoencoder at scale 1. Top row: the learned basis using a 2-layer autoencoder at a) 5th, b) 55th, and c) 299th training epoch. Bottom row: the learned basis using a multi-layer autoencoder at d) 99th, e) 199th, and f) 299th training epoch.

Representation degree is the number of self overlaps for each filter during the convolution computation. Large amount of filter overlap smooths the representation between stride locations in a convolutional layer, and low amount of filter overlap leads to abrupt changes between stride locations, creating to unnecessary edges. Therefore, representation degree controls the smoothness of the representation. However, higher representation degree requires more computation and makes the autoencoder slower and harder to converge. To strike a balance between smoothness and efficiency, in this paper we use a representation degree of 4.

Combining multiple 2-layer autoencoders with the purposed learning framework described in Section 2, we create a hierarchy of sequentially trained autoencoders that realizes DRT.

3 Experimental Results

To explore the efficacy of the proposed DRT, we evaluate our method on two datasets comprising of very different types of image content, CelebFaces Attributes [14] and Cityscapes [15]. CelebFaces consists of human face images, where each face takes up a large portion of the image. Cityscapes consists of images taken from car dash cameras while driving around cities. Each Cityscapes image contains a large variety of objects where each object takes up a small portion of the image, including buildings, roads, cars, and pedestrians. The boundaries and edges of these object contributes the large amount of high frequency information to the image.

We design two experiments to show the representation flexibility and efficiency of DRT.

In the first experiment, we demonstrate the representation flexibility by evaluating one autoencoder layer of DRT under two different parameter settings. Specifically, we compare autoencoders with two different stride size while keeping the hidden layer channel number to be 3 and keeping the representation degree θ_d to be 4. The stride size of the first autoencoder is set to be 32, and the stride size of the second autoencoder is set to be 8. The two stride size we compare are 32 and 8. The number is chosen to show large visual gap between the learned basis. Figure 5 shows the learned bases for both parameter settings. We observe that the autoencoder trained with larger stride size only characterizes the coarse structures of human face, while the autoencoder trained with smaller stride size captures finer details on the face including hair, eyes, nose, etc.. The spectrum magnitude analysis shows that the basis learned with large stride size characterizes signal information in a narrow band around 0 Hz frequency, while the basis learned with small stride size characterizes signal information in a broader band around 0 Hz frequency. In the spectrum magnitude analysis, we also see non-uniform distribution in high frequency bands for both learned bases, indicating that autoencoder is able to learn complex non-linear features that consist of both high and low frequency information. Furthermore, user can adjust the coarseness/fineness at a scale by selecting the trained representation basis from a specific training epoch, as the autoencoder progressively learns finer details and sharper edges. Figure 4 shows a learning progression of a basis. In short, the autoencoder of DRT demonstrates great representation flexibility for characterizing signal information at a user controlled detail level.

In the second experiment, we apply DRT to CelebFaces and Cityscapes. In this particular DRT, we use 3 autoencoders layers. To accommodate the large difference in information distribution of the two datasets, we use a larger stride size to focus on low frequency representation learning in CelebFaces and we use smaller stride size to focus on high frequency representation learning in Cityscapes. The stride size of the three autoencoder in DRT is 32, 16, 4 for CelebFaces, and 8, 2, 2 for Cityscapes. Figure 6 shows the learned DRT bases for CelebFaces, and Figure 7 show the learned DRT bases for Cityscapes. By visual comparison, we can see that each learned basis contains unique and significant visual characteristics of the original image. The visual quality progressively improves as the reconstructed image uses more learned basis. The spectrum magnitude analysis shows that even the two dataset possess different information distribution, DRT is able to disentangle the signal information at clustered area for both datasets and decompose the signal into compact non-linear representations.

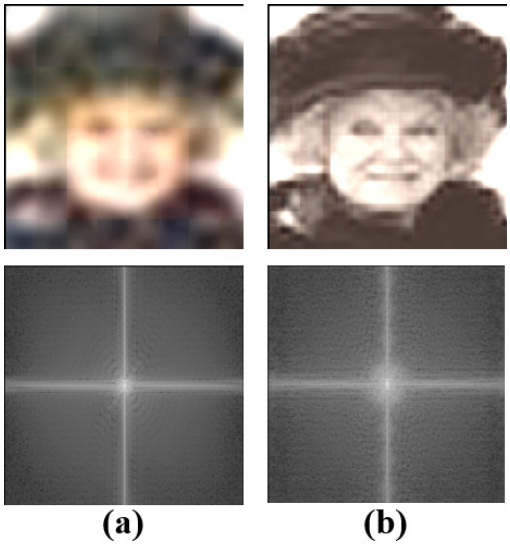

Fig. 5: Learned basis for the first decomposition scale. Top row: learned basis in time domain. Bottom: the spectrum magnitude of the learned basis. Image center corresponds to the frequency at 0 Hz. a) basis learned using an autoencoder with a stride size of 32. b) basis learned using an autoencoder with a stride size of 8.

Fig. 6: Visual results of basis learned from the CelebFaces. Top row: learned basis. Middle row: reconstructed image at scale i. Bottom row: the spectrum magnitude for the corresponding basis. Image center corresponds to the frequency at 0 Hz.

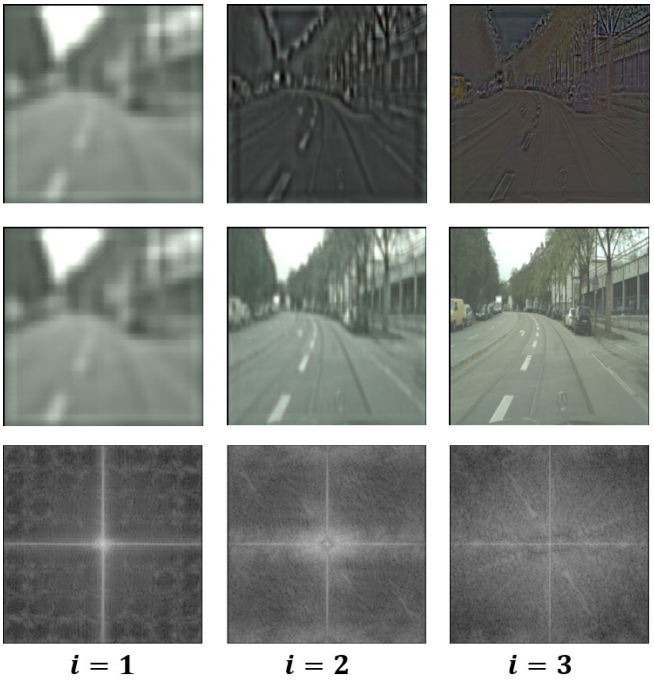

Fig. 7: Visual results of basis learned from the Cityscapes. Top row: learned basis. Middle row: reconstructed image at scale i. Bottom row: the spectrum magnitude for the corresponding basis. Image center corresponds to the frequency at 0 Hz.

4 Conclusion

In this paper, we presented Deep Residual Transform (DRT), a data-driven MID strategy where input signal is transformed into a hierarchy of independently learned non-linear representational residuals. The proposed DRT was realized by leveraging a hierarchy of sequentially trained autoencoders. Experimental results showed that the proposed DRT achieved highly efficient information decomposition on both CelebFaces and Cityscapes dataset, and demonstrated DRT's representation flexibility by showing how parameter adjustments can control the coarseness/fineness at each scale. In the future, we will investigate the use of generative models for to learn basis in DRT. We will also automate the parameter adjustments int DRT.

Acknowledgments

We thank Natural Sciences and Engineering Research Council and the Canada Research Chairs program.

References

[1] S. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transactions on Image Processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

[2] G. K. Wallace, "The jpeg still picture compression standard," vol. 34, no. 4, 1991.

[3] R. Strickland and H. Hahn, "Wavelet transform methods for object detection and recovery," *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 724–735, 1997.

[4] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[5] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[6] E. Adelson, C. Anderson, J. Bergen, P. Burt, and J. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[7] K. Subr, C. Soler, and F. Durand, “Edge-preserving multiscale image decomposition based on local extrema,” *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 1–9, 2009.

[8] E. Gastal and M. Oliveira, “Domain transform for edge-aware image and video processing,” *Proceedings of SIGGRAPH*, no. 4, 2011, Vancouver, Canada.

[9] P. Shao, S. Ding, L. Ma, Y. Wu, and Y. Wu, “Edge-preserving image decomposition via joint weighted least squares,” *Computational Visual Media*, vol. 1, no. 1, pp. 37–47, 2015.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Proceedings of the Conference on Neural Information Processing Systems*, pp. 2672–2680, December 2014, Montreal, Canada.

[11] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, December 2013.

[12] A. V. D. Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” *Proceedings of the Conference on Neural Information Processing Systems*, December 2017, Long Beach, CA.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” *Proceedings of International Conference on Computer Vision*, June 2015, Boston, MA.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, Las Vegas, NV.