Exploring images with deep learning for classification, retrieval and synthesis

Liu, Y.

Chapter 1

Introduction

1.1 Motivation

In 2018, the number of mobile phone users will reach about 4.9 billion. Assuming an average of 5 photos taken per day using the built-in cameras would result in about 9 trillion photos annually. In addition, these photos are frequently uploaded, shared and retrieved in social networks and thus have become an important part of our daily lives. However, it is challenging to mine semantically meaningful visual information from such a huge amount of data. Thanks to the major advances of deep neural networks since 2012, they have been a powerful tool to help analyze visual content for a variety of tasks and have triggered a massive amount of research in content based multimedia analysis and computer vision. This thesis aims towards *developing new paradigms and architectures in deep learning* to address three common and important research themes: classification, retrieval and synthesis. As shown in Figure 1.1, we visibly depicts the three themes.

- **Classification** is the most fundamental task in the field of computer vision. It aims to correctly predict the class label for a given image, for example, we can use a classification model to classify the input horse image. In addition to image-level classification, we also study the tasks of pixel-level classification, including semantic segmentation and edge detection. (Chapters 2 and 3)
- **Retrieval** aims to efficiently search for similar samples from the database to the query. For instance, we develop a retrieval model to retrieve similar horse images. Besides, we also consider the cross-modal retrieval problem between images and texts, and do some work to bridge the modality gap between vision and language. (Chapters 4, 5, 6 and 7)
- **Synthesis** is able to generate new image samples that never existed in the image database. For example, by training a synthesis model, we can translate a horse image to a zebra image. In addition, we can synthesize diverse zebra images based on different branches of the network. In this thesis, we mainly focus on two synthesis applications: image-to-image translation and fashion style transfer. (Chapter 8)

In the next sections, we first introduce the background and developments related to the three themes in recent years. Then we present the thesis outline, our research questions and main contributions.

1.2 Background and Related Work

Deep learning [1, 2] has been one of the pillars of numerous artificial intelligence research fields, such as computer vision, machine learning and natural language

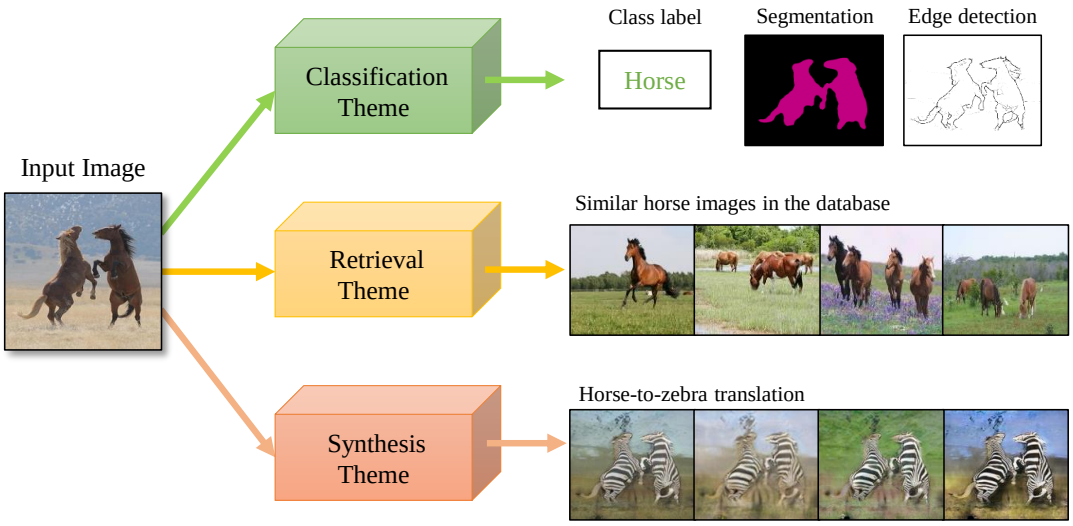


Figure 1.1: Conceptual illustration of the three research themes in this thesis, including classification, retrieval and synthesis.

processing. By distilling high-level semantic information with deep network architectures, deep learning approaches can help narrow the gap between visual representations and human-level vision. In recent years, deep learning has been extensively studied in the field of computer vision to help tackle many challenging tasks, such as image classification, image retrieval and image synthesis.

1.2.1 Classification

In recent decades, exploiting and developing convolutional neural networks (CNNs) [3] has been a leading and promising trend in computer vision community. CNNs can explore high-level visual concepts in images by employing deep architectures composed of multiple neural layers. In 2012, Krizhevsky *et al.* [4] proposed a new CNN model named AlexNet for generic image classification, which has been a milestone in the developments of CNNs. Its success in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competitions [5] motivates a huge amount of work leveraging CNNs to solve various vision tasks. According to the empirical observations in [6], CNNs based approaches can achieve new state-of-the-art performance for different recognition tasks by fine-tuning the ImageNet models on the target datasets. We summarize the related work on classification from the following four aspects.

Increasing the depth

A dominant line of research on CNNs is to increase the network depth to learn more discriminative representations (Figure 1.2(a)). For instance, the depth has increased

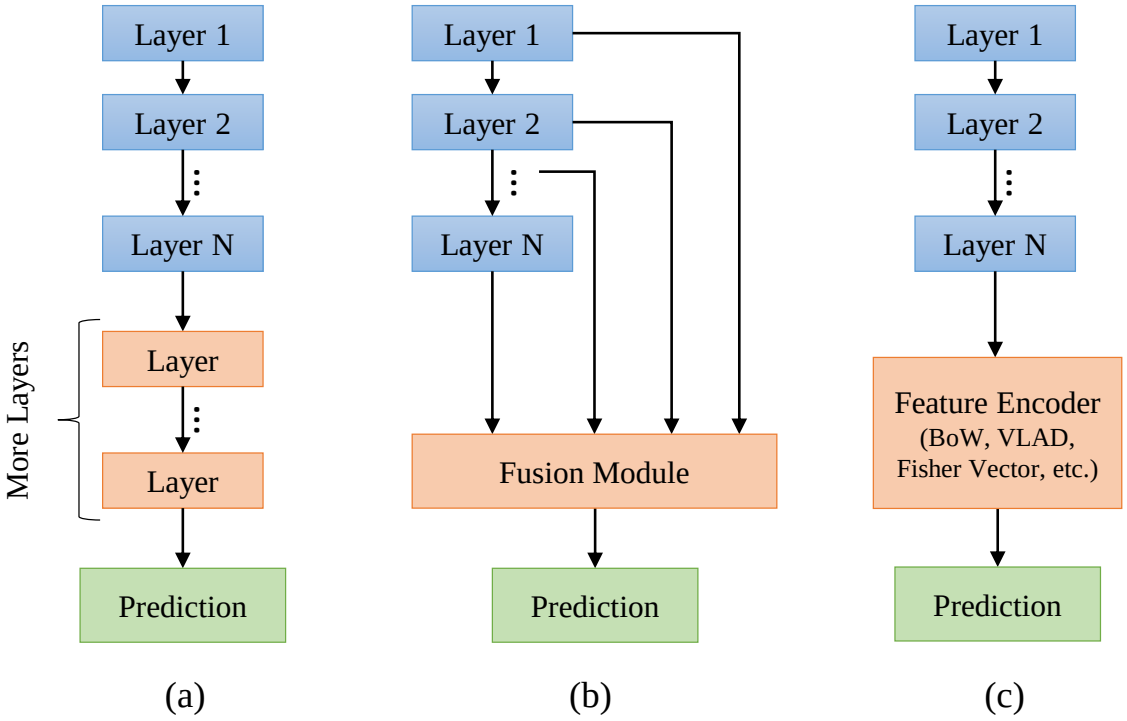


Figure 1.2: Illustration of three classification pipelines. (a) increasing the depth with more new layers. (b) fusing intermediate layers to produce an aggregation feature. (c) encoding deep features with sophisticated feature encoders.

from several layers (*e.g.* LeNet [3] and Alexnet [4]) to several tens of layers (*e.g.* VGGnet [7] and GoogLeNet [8]). However, training deeper networks becomes more difficult because of vanishing gradients and degradation. To overcome this challenge, Highway networks [9] and ResNet [10] proposed to add shortcut connections between neighboring layers, which can help alleviate the vanishing gradient issue and ease the training convergence. Their approaches have promoted the study on constructing deeper neural networks (*e.g.* hundreds of layers) and breaking the potential bottleneck that may limit the learning capabilities. Furthermore, extended studies [11, 12, 13, 14] based on ResNet provided additional insights by delving into the residual learning mechanism. Nevertheless, it is non-tractable to optimize much deeper neural networks due to the large amount of network parameters and the expensive cost of physical memory.

Fusing multiple layers

An alternative to creating shortcuts between adjacent layers is to integrate existing intermediate layers in a deep neural network to generate a fused feature (Figure 1.2(b)), rather than deepening the network with additional new layers. Commonly, the topmost activations in deep networks (*i.e.* fully-connected layers) can act as the most important features to describe the image content. However, it

is important to note that intermediate activations (*i.e.* convolutional layers) can also provide informative and complementary clues about images, including low-level boundaries, textures and spatial contexts. Therefore, researchers [15, 16] began to transfer their attention to intermediate layers, and explored their influence on the classification performance. In contrast to using pre-trained models, extensive research efforts [17, 18] turned to training deep fusion networks where multi-level intermediate layers are fused together by adding new side branches. It is worth noting that the fused information occurs not just from adjacent layers but from the earliest layers as well. In the literature, deep fused representations have been shown to generate better predictions due to integrating the strengths of different intermediate layers within deep neural network.

Encoding deep features

Although CNNs are able to express more powerful visual features, they have weak robustness to severe geometrical deformations and spatial contexts. Fortunately, sophisticated encoding techniques including BoW [19], VLAD [20] and Fisher Vector [21] have been adopted to address these issues. Motivated by the strengths of encoding techniques, it is natural to encode deep features to further improve their discriminatory power (Figure 1.2(c)). To obtain local features from CNN models, most approaches [22, 23, 24, 25] have examined local patches or region proposals in one image. The local CNN features are used to construct a visual codebook, based on which an encoder technique can be used to aggregate them to a deep image representation. For example, Gong *et al.* [22] employed image patches at multiple scales, and then aggregated local patch responses at the finer scales with the VLAD method. Yoo *et al.* [25] utilized multi-scale dense local CNN features to compute the Fisher Vector kernels.

Pixel-level classification

In addition to image-level classification, CNNs also show strong generalization power for diverse tasks of pixel-level classification, such as semantic segmentation [26, 27, 28], edge detection [29, 30, 31], depth estimation [32, 33, 34] and saliency detection [35, 36, 37]. In particular, fully convolutional networks (FCNs) [26] have become a fundamental architecture to perform pixel-level predictions. Specifically, FCNs are recast from the pre-trained CNN counterparts, by replacing the fully-connected layers with extra convolutional layers while retaining the parameters. In this way, the size of the input images can be arbitrary and the output can be viewed as two-dimensional feature maps. In addition, it is beneficial to extract richer region features from FCNs, compared to a global representation from CNNs.

1.2.2 Retrieval

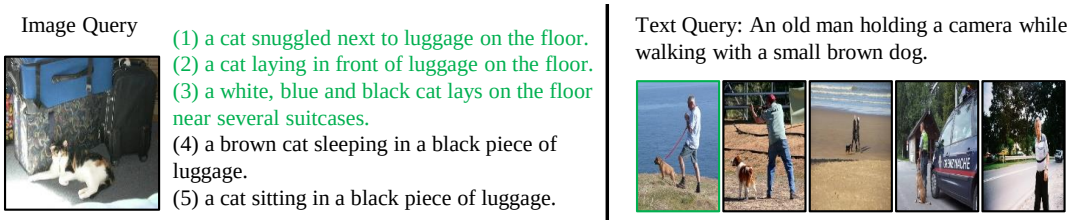
One of the primary aims of image retrieval is to search for similar images (usually based on pictorial content) to the query from the database. It has become important to numerous practical scenarios (*e.g.* Google image search, face recognition, *etc.*) and therefore has triggered a massive amount of research activities in both multimedia and computer vision fields [19, 38, 39]. Bag-of-Words (BoW) is one of the most widely-used models in image retrieval systems, where local features, such as SIFT [40] and color clues [41], are quantized to visual words based on a pre-trained codebook. Then, similar to document retrieval [19, 39], an inverted index structure is built with the visual words towards making the retrieval system scalable and efficient. However, image retrieval remains challenging in bridging low-level image representations and high-level semantic concepts.

Image retrieval

To alleviate the above challenge, recent works in the literature have paid attention on utilizing deep visual features for image retrieval [42, 43, 44]. The work of Wan *et al.* [42] suggested that a deep CNN model pre-trained on a large dataset can be transferred to new content-based image retrieval (CBIR) tasks and fine-tuning the model with a similarity metric could further boost the retrieval performance. Babenko *et al.* [45] focused on holistic descriptors where the whole image was mapped to a single deep feature vector. They further designed a simple global image descriptor based on sum-pooled convolutional features for image retrieval. Zheng *et al.* [46] proposed a deep embedding method using deep features as global and regional signatures instead of a Hamming embedding[47]. It is an incorporation of the SIFT descriptor and CNN features and could achieve promising improvements. Moreover, Zheng *et al.* [48] presented a comprehensive review on SIFT and CNN-based methods and discussed the benefits of integrating SIFT and CNN features.

Cross-modal retrieval

Nowadays, multimedia data in various media types (*e.g.* image, video, text, and audio) is growing exponentially due to the increasing popularity of the Internet and social networks. This trend motivates a massive amount of research activities in multi-modal understanding and reasoning. For example, we can recognize a picture of a panda after hearing the description "black and white bears" without ever having seen one. This demonstrates the cross-modal interaction between vision and language. These heterogeneous data offers us the opportunity to understand the world better, while giving rise to the challenges of bridging different modalities. Specifically, the matching problem between images and texts [49, 50, 51, 52, 53, 54] is one of the most important tasks in multi-modal research. In practice, image-text matching



(a) Image-to-text retrieval

(b) Text-to-image retrieval

Figure 1.3: Example of cross-modal retrieval. (a) Given an image query, related text samples are retrieved to describe the image. (b) For a text query, it can search for several image samples from the database. The matched samples are highlighted with green color.

approaches are usually developed for cross-modal retrieval (Figure 1.3). This task remains challenging due to the heterogenous representations and the cross-modal gap between vision and language, which is also a core issue for other multi-modal applications such as image captioning [55, 56] and visual question answering [57, 58], zero-shot recognition [59, 60].

With the increasing progress of deep learning, research efforts have been made to incorporate Canonical Correlation Analysis (CCA) [61] into deep neural networks [49, 50, 51, 62, 63]. However, existing deep CCA models rely on expensive decorrelation computations, which limit their generalization abilities at large-scale data. Alternatively, a number of recent approaches [52, 55, 64, 65, 66] address the task by designing two-branch networks to embed visual and textual features into a common latent space, and then learn latent embeddings by optimizing a ranking loss to discriminate matched and unmatched image-text pairs. For instance, Wang *et al.* [53] built a simple and efficient matching network to preserve the structure relations between images and texts in the latent space. To associate image regions with words, the attention mechanism was integrated into visual-textual embedding models [67, 68]. In addition to the pairwise ranking loss, recent approaches [69, 70] leveraged extra loss functions (*e.g.* instance loss and classification loss) to enhance the discrimination of the learned embedding features.

Another line of research [71, 72, 73, 74, 75] focused on learning dual embeddings between two modalities, *e.g.* projecting visual features into the textual feature space and vice versa. For instance, Feng *et al.* [71] proposed a correspondence cross-modal autoencoder model. 2WayNet [76] built the projections between two modalities and regularized them with Euclidean loss. Recently, Gu *et al.* [77] utilized two generative models to synthesize grounded visual and textual representations. Essentially, these dual embedding models are motivated by autoencoders.

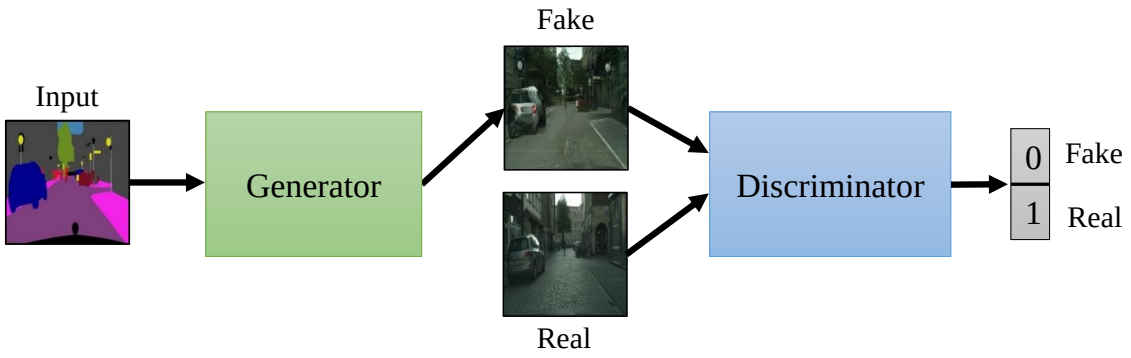


Figure 1.4: Illustration of the GAN framework. In this example, given a labelled map, the generator can synthesize a fake photo image similar to the real one, but The discriminator learns to correctly classify real and fake images.

1.2.3 Synthesis

Together with the increasing progress of deep neural networks, numerous approaches based on supervised learning have been developed to address diverse image translation tasks, such as contour detection [31], semantic segmentation [26] and face conversion [78]. However, these supervised models highly depend on a large amount of fully labelled image pairs which are time consuming to create manually and sometimes biased when collecting annotated data. In addition, some ground-truth data are not available in some cases, for example, the painting stylization transfer between Monet to Van Gogh. Driven by these limitations, researchers have turned to examine unsupervised learning approaches to break the bottleneck of limited data.

Generative adversarial networks

Generative models have attracted increasing attention with the emergence of generative adversarial networks (GANs)[79]. Informally, the GAN framework can be viewed as a game between two players: the generator and the discriminator (Figure 1.4). To be specific, the generator aims to synthesize fake images and tries to trick the discriminator into thinking that the synthesized images are real. In contrast, the discriminator needs to distinguish the real images from the fake images. By continuing this game iteratively, both players learn to become better until the generator can generate realistic-looking images and the discriminator can not tell real and fake samples. Typically, a simple analogy is that: an art forger (the generator) attempts to forge artistic paintings, but an art investigator (the discriminator) is able to detect imitations. In recent years, GANs have been widely adopted for addressing a wide range of image synthesis applications, such as style transfer [80, 81], texture synthesis [82, 83] and text-to-image synthesis [84, 85]. To improve the quality and diversity of generative models, conditional GANs (cGAN) have been designed to

guide image generation conditioned on class labels [86], attributes [87], images [88] and texts [84].

Image-to-image translation

GANs have shown great success on the task of general-purpose image-to-image translation [88], which learns to model mapping functions between different image domains. Many recent approaches [89, 90, 91, 92] were focused on using unpaired images to tackle the problem of unsupervised image translation. In addition to the adversarial constraint, they further exploited extra constraints to enhance relations of two different domains. On the one hand, some of them [89, 90, 93] fed image samples into a unified encoder to discover their common representations. Then another generator was used to translate common representations to samples in the target domain. On the other hand, some work [92, 94] attempted to relate two different domains by using additional self-constraints within one domain. Representatively, CycleGAN [94] proposed a cycle-consistency constraint that can reconstruct the input image itself.

Fashion style transfer

Online shopping has driven a range of fashion oriented applications recently, for example, fashion clothing retrieval [95], fashion recommendation [96], fashion parsing [97] and fashion style transfer [81]. Specifically, fashion clothing swapping, which is a common application belonging to fashion style transfer, aims to visualize what the person would look like with the target clothes. This application allows consumers to see what they would look like by wearing different clothes, without the effort of dressing them physically. In the past, this problem has been studied in the fields of multimedia and computer graphics [98, 99, 100, 101]. For example, the work in [102] used an image-based visual hull rendering approach to transfer the appearance of a target garment to another person image. The ClothCap approach [103] captured the 3D deformations of the clothing and estimated the minimally clothed body shape and pose under the clothing. These non-parametric solutions [100, 104] involve using extra information to model the deformations, such as from motion capture, 3D measurements and depth sensors. During the test stage, they still require online image warping or registration algorithms which are time-consuming for real-time applications. Recent research turned to address this problem using deep generative approaches (*e.g.* GANs), without requiring complicated 2D image warping and 3D graphic algorithms. For example, FashionGAN [85] employed a textual description as condition to perform the clothing swapping. The methods in [105, 106] took a stand-alone and flat clothing image to re-dress the person in the reference image.

1.3 Thesis Outline and Research Questions

In Section 1.2, we have introduced recent advances on the three research themes. Although deep learning is leading state-of-the-art performance for numerous tasks, we should notice its limitations and challenges, such as theoretical interpretability, model complexity, training with limited data, *etc.* There is still considerable space for promoting the developments of deep learning. In the next research chapters (Chapters 2-8), we propose new approaches to address the research questions (**RQ**) and challenges in terms of the three research themes. In Chapter 9, we discuss our main findings, limitations & possible solutions and future research directions.

- **Chapter 2** aims to address the first research question **RQ 1: How can we develop a simple and efficient deep fusion network upon a plain CNN?** As discussed in Section 1.2.1, some works [17, 18, 26, 31] attempt to create new side branches upon a plain CNN and integrate multi-level intermediate layers to generate a fused representation. However, they still have two main limitations. First, some of them spend a large number of new parameters creating the side branches. For example, DAG-CNNs [18] add several fully-connected layers on top of intermediate convolutional layers, which will largely increase the total number of parameters. Second, the fusion modules for integrating different side branches are inferior. DAG-CNNs [18] and FCN-8s [26] use a simple sum pooling to fuse the side branches, which fails to consider the weights of different side branches. Although HED [31] employs a 1×1 convolution to learn the fused weights, they are shared over spatial dimensions, failing to discover the spatial properties. In this chapter, we propose a novel convolutional fusion network (CFN) built on top of plain CNNs, which can aggregate intermediate layers with adaptive weights and generate a discriminatively fused representation. This chapter is based on the published papers [107, 108]:

  > **Liu, Y.**, Guo, Y., and Lew, M.S., "On the Exploration of Convolutional Fusion Networks for Visual Recognition." Proceedings of the 23rd International Conference on MultiMedia Modeling (MMM), 2017. **(Best Paper Award)**
  >
  > **Liu, Y.**, Guo, Y., Georgiou, T., and Lew, M.S., "Fusion that matters: convolutional fusion networks for visual recognition." Multimedia Tools and Applications, 2018.

- The work in **Chapter 3** aims to tackle the second question **RQ 2: How can we explore diverse supervision that can adapt to different intermediate layers in deep neural networks for robust edge detection?** Edge detection that aims to distinguish important edges from image pixels,

can generally act as a fundamental task for other high-level vision applications, like object detection and segmentation. Recently, the developments in the design of edge features have moved from carefully-engineered descriptors [109, 110, 111, 112] to hierarchical deep features [29, 30, 31, 113]. Nevertheless, we should still realize one difficult issue in edge detection that is caused by *false positives*: many non-edge pixels are incorrectly predicted as edges when comparing with the human annotated ground-truth. To correct the false positives earlier, HED [31] imposes the ground-truth supervision on the intermediate layers while training the deep model. However, using only a general supervision (*i.e.* the ground-truth annotation) for all the layers is inconsistent with the diverse representations of hierarchical layers. In addition, the general supervision can not be well-suited to all intermediate layers. In contrast to using the general supervision, we propose and develop relaxed deep supervision (RDS) within convolutional neural networks for robust edge detection. This chapter is based on the published paper [114]:

> **Liu, Y.** and Lew, M.S., "Learning Relaxed Deep Supervision for Better Edge Detection." Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

- In **Chapter 4**, we move our attention to the retrieval theme and tackle the third question **RQ 3: How can we incorporate deep visual representations into the inverted index structure for accurate and efficient image retrieval?** A robust image retrieval system should be typically optimized regarding two factors: accuracy and efficiency. To increase the retrieval accuracy, some works [22, 42, 45] begin to utilize deep visual features to discover the similarities among images. However, they are inefficient due to relying on the nearest neighbouring search. To maintain the efficiency, traditional methods [19, 39], take advantage of the inverted index structure that is able to reduce computational time and memory cost for scalable image search. Regarding both the accuracy and efficiency, we exploit a DeepIndex framework for accurate and efficient image retrieval, by incorporating deep visual features into the inverted index scheme. This chapter is based on the published paper [115]:

  > **Liu, Y.**, Guo, Y., Wu, S., and Lew, M.S., "DeepIndex for Accurate and Efficient Image Retrieval." Proceedings of the 5th ACM on International Conference on Multimedia Retrieval (ICMR), 2015.

- In addition to image retrieval, in **Chapter 5** we further address the problem of cross-modal retrieval with **RQ 4: How can we build a deep matching network to unify images and texts into a more discriminative space without increasing the number of network parameters?** The image-text matching problem remains challenging due to the heterogenous representations and the cross-modal gap between two modalities. In recent

years, a variety of multi-modal deep neural networks have been proposed to model the matching task [52, 53, 76]. However, the multi-modal matching performance is still far from competitive with the intra-modal tasks, for example, image retrieval. In this chapter, we introduce an efficient approach to couple visual and textual features based on a new recurrent residual fusion (RRF) building block. This chapter is based on the published paper [64]:

> **Liu, Y.**, Guo, Y., Bakker, E.M., and Lew, M.S., "Learning a Recurrent Residual Fusion Network for Multimodal Matching." Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), 2017.

- In terms of cross-modal retrieval, in **Chapter 6** we further pose the fifth research question **RQ 5: How can we preserve both inter-modal correlations and intra-modal consistency for learning robust visual and textual embeddings?** Currently, there are two main paradigms to perform visual-textual embeddings. The first one is to learn a common space where related images and texts can be unified into similar latent embeddings [52, 53, 76]. Second, it exploits dual embeddings by reconstructing an input feature in the source space to be the one in the target space [72, 76, 77]. Both the latent and dual embeddings can capture inter-modal semantic correlations between visual and textual data. In addition, they can be combined together to integrate individual advantages. However, they fail to preserve the intra-modal semantic consistency, *i.e.* image-to-image and text-to-text. Importantly, a robust embedding method should be able to reconstruct representations of both the source and target modalities. To achieve this, we propose cycle-consistent embeddings in a deep neural network for matching visual and textual representations. This chapter is based on the submitted journal paper [116]:

  > **Liu, Y.**, Guo, Y., Liu, L., Bakker, E.M., and Lew, M.S., "CycleMatch: A Cycle-consistent Embedding Network for Image-Text Matching." Submitted to IEEE Transactions on Multimedia (In Revision).

- In **Chapter 7**, we aim to integrate both matching and classification by answering the sixth question **RQ 6: How can we design a unified network for joint multi-modal matching and classification?** We note that, learning visual-textual embeddings is influenced by the notable variance in images or texts. For example, in the MSCOCO dataset [117], each image is described with five sentences from human labelers. Although the sentences can consistently mention some primary objects in the image, they have some biased differences that may make it difficult to perform a robust matching. However, object labels can generally provide more consistent and less biased information than sentences. Classification with the object labels is beneficial

to correct the biased sentences and improve the image-text matching. Additionally, the matching component can help the classification component to generate a discriminative multi-modal representation. Unlike many current approaches which only focus on either multi-modal matching or classification, we propose a unified network to jointly learn Multi-modal Matching and Classification (MMC-Net) between images and texts. This chapter is based on the submitted paper [118]:

> **Liu, Y.**, Liu, L., Guo, Y., and Lew, M.S., "Learning Visual and Textual Representations for Multimodal Matching and Classification." Pattern Recognition, vol 84: 51-67, 2018.

- **Chapter 8** presents two applications about image synthesis: image-to-image translation and fashion style transfer.

  For image-to-image translation, we pose the seventh research question **RQ 7: What factors will affect the performance of generative models on the translation tasks?** Image-to-image translation between different domains is a common image synthesis task, with the aim of arbitrarily manipulating the source image content given a target one. To tackle the challenging case of unpaired image-to-image translation, CycleGAN [94] presents a cycle-consistency loss by reconstructing the generated image back to the source domain. In conjunction with the original adversarial loss, the cycle-consistency loss is beneficial to constrain the unsupervised domain mappings. CycleGAN has become a fundamental approach for general-purpose image-to-image translation, while few work investigate the important factors within it. To address the problem, we present an extensive and empirical study on cycle-consistent generative networks. This work is based on the published paper [119]:

  > **Liu, Y.**, Guo, Y., Chen, W., and Lew, M.S., "An Extensive Study of Cycle-Consistent Generative Networks for Image-to-Image Translation." Proceedings of the 24th International Conference on Pattern Recognition (ICPR), 2018.

  In terms of fashion style transfer, we need to tackle the last research question **RQ 8: How can we exploit a generative model to directly transfer the fashion style between two person images?** Currently, fashion style transfer based on image synthesis has become a popular application for online shopping. Specifically, fashion clothing swapping aims to visualize what the person would look like with the target clothes. It can be viewed as a specific task belonging to fashion style transfer. Recently, FashionGAN [85] specifies a textual description and uses it to re-dress the person in the reference image. other works in CAGAN [105] and VITON [106] employ a stand-alone and flat clothing image to condition the image synthesis, which may provide richer visual content than the textual description. However, the stand-alone and flat

clothing images are not always available to users or consumers. Therefore, we pose a more practical task, that is, person-to-person clothing swapping, where the input condition is also a person image like the reference image. In this case, the goal becomes transferring the wearing clothes between two person images. It is more challenging as the desired clothes worn on the condition person image have varying deformations due to different human poses. To tackle this challenge, we propose a novel multi-stage generative network (SwapGAN) that integrates three generators to perform a multi-stage synthesis process. This work is based on the submitted journal paper [120]:

> **Liu, Y.**, Chen, W., Liu, L., and Lew, M.S., "SwapGAN: A Multi-stage Generative Approach for Person-to-Person Fashion Style Transfer." Submitted to IEEE Transactions on Multimedia (In Review).

- Finally, **Chapter 9** summaries the main findings from the research of this thesis. Also, we discuss the limitations and potential solutions, and point out directions for future research.

Additionally, this thesis draws on insights and experiences from the related work in other publications during my PhD studies:

- **Liu, Y.** and Lew, M.S., "Improving the Discrimination between Foreground and Background for Semantic Segmentation." Proceedings of the 24th IEEE International Conference on Image Processing (ICIP), 2017.
- **Liu, Y.**, Guo, Y., and Lew, M.S., "What Convnets Make for Image Captioning." Proceedings of the 23rd International Conference on MultiMedia Modeling (MMM), 2017.
- Guo, Y., **Liu, Y.**, Lao, S., Bakker, E.M., Bai, L., and Lew, M.S., "Bag of Surrogate Parts Feature for Visual Recognition." IEEE Transactions on Multimedia, vol 20: 1525-1536, 2018.
- Shan, H., **Liu, Y.**, and Stefanov, T., "A Simple Convolutional Neural Network for Accurate P300 Detection and Character Spelling in Brain Computer Interface." International Joint Conference on Artificial Intelligence (IJCAI), 2018.
- Guo, Y., **Liu, Y.**, de Boer, M.H.T., Liu, L., and Lew, M.S., "A Dual Prediction Network for Image Captioning." Proceedings of the 19-th IEEE International Conference on Multimedia and Expo (ICME), 2018.
- Georgiou, T., Schmitt, S., Olhofer, M., **Liu, Y.**, Back, T., and Lew, M.S., "Learning Fluid Flows." Proceedings of International Joint Conference on Neural Networks (IJCNN), 2018.

- Guo, Y., **Liu, Y.**, Bakker, E.M., Guo, Y., and Lew, M.S., "CNN-RNN: a large-scale hierarchical image classification framework." Multimedia Tools and Applications, vol 77: 10251-10271, 2018.
- Guo, Y., **Liu, Y.**, Georgiou, T., and Lew, M.S., "A review of semantic segmentation using deep neural networks." International Journal of Multimedia Information Retrieval, vol 7: 87-93, 2018.
- Jia, Q., Fan, X., **Liu, Y.**, Luo, Z., and Guo, H., "Hierarchical projective invariant contexts for shape recognition." Pattern Recognition, vol 52: 358-374, 2016.
- Guo, Y., **Liu, Y.**, Oerlemans, A., Lao, S., Wu, S., and Lew, M.S., "Deep learning for visual understanding: A review." Neurocomputing, vol 187: 27-48, 2016.
- Guo, Y., Lao, S., **Liu, Y.**, Bai, L., Liu, S., and Lew, M.S., "Convolutional Neural Networks Features: Principal Pyramidal Convolution." Proceedings of the 16th Pacific-Rim Conference on Multimedia (PCM), 2015.

1.4 Main Contributions

The research of this thesis contributes at three levels: models and algorithms, practical scenarios and empirical analysis.

1.4.1 Models and algorithms

From Chapter 2 to Chapter 8, we develop new approaches based on deep learning to address the research questions regarding the three themes. The key contributions in these approaches are listed below.

An efficient deep fusion model for image classification. We propose a novel deep fusion network, the convolutional fusion network (CFN), where we can efficiently integrate multiple intermediate layers in CNNs with adaptive weights. In addition, our CFN is adaptive to not only image-level classification, but also pixel-level classification.

A diverse deep supervision algorithm for edge detection. In contrast to prior work using a general supervision, we develop relaxed deep supervision (RDS) with additional relaxed labels. Consequently, more discriminative layers can process more false positives in edge detection. RDS can incorporate the diversities into the supervisory signals to improve the performance of edge detection.

An accurate and efficient model for image retrieval. We propose a novel image retrieval approach (DeepIndex) which can integrate deep visual representations

with the inverted index scheme. Our approach takes advantage of the discriminatory capabilities of deep features and the efficient search of the inverted index.

An efficient image-text matching model for cross-modal retrieval. We develop a novel recurrent residual fusion network (RRF-Net) to couple visual and textual features. Since RRF-Net connects the residual learning with the recurrent mechanism, it can recursively improve visual-textual embeddings while sharing the network parameters. In addition, we develop a fusion module to efficiently integrate intermediate recurrent outputs.

A cycle-consistent embedding algorithm for cross-modal retrieval. To preserve both inter-modal correlations and intra-modal consistency, we propose cycle-consistent embeddings by cascading dual mappings and reconstructed mappings in a cyclic fashion. Our embedding method can effectively promote the performance of cross-modal retrieval, compared to traditional embedding methods.

A unified model for multi-modal matching and classification. We propose a unified network (MMC-Net) to jointly model multi-modal matching and classification. Our approach can suggest that combining the matching and classification components can help boost each other. In addition, we employ a multi-stage training algorithm to make the two components compatible.

Two extended deep generative models for image-to-image translation. As few work investigate the important factors within cycle-consistent generative networks (CycleGAN), we present two extended models, namely Long CycleGAN and Nest CycleGAN, and then conduct an extensive and empirical study on the models. Our work examines the benefits of using more generators and cycles on the generation quality.

A multi-stage generative model for fashion style transfer. In contrast to traditional non-parametric approaches, we propose a novel multi-stage generative network (SwapGAN) to transfer the clothing style in one person image to another one. The SwapGAN model can be trained end-to-end with three different generators and one discriminator.

1.4.2 Practical scenarios

In addition to improve the performance of diverse tasks, our research also aims towards adapting to practical scenarios in real world.

Accurate and efficient image retrieval. Image retrieval is a widely used application in ours lives. In some cases, the retrieval speed is the same important as the accuracy. Our DeepIndex framework (in Chapter 4) can take into account both accuracy and efficiency in image retrieval. Specifically, deep visual features can help

improve the retrieval accuracy, and the inverted index scheme is more efficient than the nearest neighboring search.

Joint multi-modal matching and classification. In practice, we may need to search for similar samples as the query sample, and know its class label as well. Motivated by this need, we develop MMC-Net (in Chapter 7) to unify both multi-modal matching and classification in one model, unlike prior approaches which focus on either matching or classification. Our work shows a simple and efficient way to fulfill the two tasks simultaneously.

Person-to-person fashion style transfer. Prior work performs the clothes-to-person style transfer, however, in practice stand-alone and flat clothing images are not always available. Instead, our work (in Chapter 8) aims to address a more practical case, in which the goal is to swap the clothes between two person images directly. This task becomes more challenging due to varying human poses. Our SwapGAN is proposed to solve this practical problem by cascading three generators in a multi-stage manner.

1.4.3 Empirical analysis

Furthermore, this thesis provides numerous experiments and in-depth analysis, which can help motivate further research on the three research themes.

Fusion that matters deep neural networks. Instead of deepening neural networks with more layers, our CFN model (in Chapter 2) is an efficient alternative to improving the capabilities of CNNs while maintaining the model complexity. In the experiments, we provide a detailed analysis to verify the effectiveness of CFN. In addition, we compare CFN with other deep models and offer an extensive discussion about them. Moreover, our CFN can be adaptive to different computer vision tasks like image classification, semantic segmentation, edge detection, *etc.* In a nutshell, our work suggests that deep fusion networks can efficiently promote the feature representational abilities of plain CNNs.

Cycle-consistency that matters visual-textual embeddings. Our proposed cycle-consistent embedding approach (in Chapter 6) is an integration of three embeddings, namely dual embedding, reconstructed embedding and latent embedding. Our approach can model both inter-modal correlations and intra-modal consistency while matching visual and textual representations. In the experiments, we conduct a comparable analysis between our approach and existing embedding approaches. The superiority of our approach over others can increase the awareness of using cycle-consistency for multi-modal research tasks.

Two factors that matter cycle-consistent adversarial networks. To provide deep insights into CycleGAN, we developed two extended models (in Chapter 8) for examining two factors: the number of generators and the number of cycles.

The qualitative and quantitative results for a range of translation tasks verify the benefits of using more generators and cycles, compared to the vanilla CycleGAN. The results in our study can help ease future research based on cycle-consistent generative networks.