




Letter

Heterogeneous Image Knowledge Driven Visual Perception

Lan Yan , Member, IEEE, Wenbo Zheng , Member, IEEE, and Fei-Yue Wang , Fellow, IEEE

Dear Editor,

This letter is concerned with visual perception closely related to heterogeneous images. Facing the huge challenge brought by different image modalities, we propose a visual perception framework based on heterogeneous image knowledge, i.e., the domain knowledge associated with specific vision tasks, to better address the corresponding visual perception problems. In particular, we first introduce the motivation and key idea of our framework. Then the proposed framework is described in detail. Furthermore, case studies demonstrate that the proposed framework can be successfully applied to practical tasks and provide a reference for constructing a heterogeneous image perception model based on the framework.

The development of imaging devices and the improvement of computer data processing capability provide hardware support for the real-time acquisition, transmission and computing of heterogeneous images captured from different sensors or different environments. Heterogeneous images are increasingly appearing in peoples' daily life. For instance, it is common to use typical heterogeneous images such as sketches and cartoons as avatars for personal homepages. Due to the rapid development of camera hardware and photography software, especially the popularity of cameras in smart phones in recent years, photography has become more and more convenient and frequent. Meanwhile, as displayed in Fig. 1, a growing body of people are willing to share their lives with friends on social media platforms by showing photos they have taken, sketches they have drawn, etc. Heterogeneous image has become a part of people's daily life and an important information carrier in modern human life. Therefore, faced with such a large amount of heterogeneous image data, how to process it with computer vision model and obtain the information required for application in specific scenarios has become an increasingly important demand.

However, as shown in Fig. 1, there are often large inter-domain modal differences between a heterogeneous image pair. For example, there are texture and shape differences between face caricatures and face photographs, and extreme texture variation between sketches and photos. Therefore, the classical computer vision algorithm based on natural scene images does not perform well in the processing of heterogeneous images. So, how to overcome the aforementioned problem and design an effective visual perception model to extract the desired information from heterogeneous images?

Considering that no matter how great the inter-domain modal differences of a heterogeneous image pair are, the categories contained and the semantics expressed in this heterogeneous image pair remain

Corresponding author: Lan Yan.

Citation: L. Yan, W. Zheng, and F.-Y. Wang, "Heterogeneous image knowledge driven visual perception," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 1, pp. 255–257, Jan. 2024.

L. Yan is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: ylan@hnu.edu.cn).

W. Zheng is with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China (e-mail: zwb2022@whut.edu.cn).

F.-Y. Wang is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiyue.wang@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JAS.2023.123435

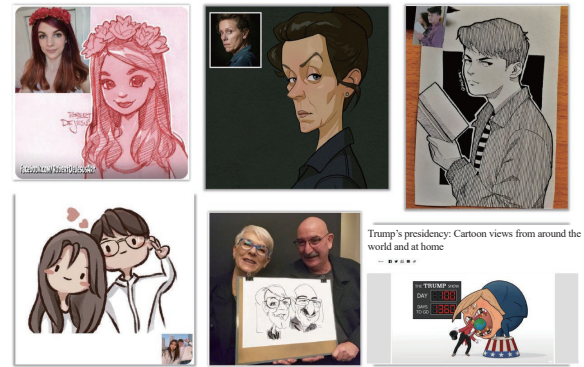


Fig. 1. Examples of people sharing heterogeneous images on social media platforms (Facebook, Instagram, Weibo and Twitter).

the same. In other words, each heterogeneous image pair has the same ontology. Therefore, different representations of one ontology have the same cognition, from this point, the sense of which can be regarded as the special domain knowledge. While the ease with which people can perceive and understand heterogeneous images and process related tasks, stems from human cognition or knowledge. So, why not introduce helpful specific domain knowledge into the visual perception model for heterogeneous images? Further, how to improve the performance of a heterogeneous image visual perception model with the introduced specific domain knowledge?

In order to answer the above questions, this letter attempts to investigate the use of specific domain knowledge of the relevant visual tasks, i.e., heterogeneous image knowledge, to deal with the corresponding visual perception problems from the characteristics of heterogeneous images, and proposes a new framework of visual perception based on heterogeneous image knowledge. Fig. 2 illustrates the key idea of our framework. As can be seen in Fig. 2, the natural image scene shows a photo of Trump, while the specific image scene presents a caricature of Trump, and the two form a group of heterogeneous image pairs. Natural images in this letter mainly refer to images that are similar or identical to those observed by the human eye under sufficient light conditions, such as portrait photos and street images taken clearly by mobile phones or cameras in a sunny day. Classical computer vision models [1], [2] are designed for tasks in natural image scenes. Due to the differences in texture and shape between caricatures and photographs, classical computer vision models can not perform well for visual tasks in specific image scenes. Therefore, unlike the general framework [3], the proposed specific framework introduces heterogeneous image knowledge into classical computer vision models through a suitable knowledge guidance mechanism to meet the needs of specific image scene tasks. Specifically, heterogeneous image knowledge refers to domain knowledge that is based on the characteristics of a particular heterogeneous image and helps to accomplish the visual task for that heterogeneous image. Knowledge is derived from information, but unlike information contains a lot of noise. Knowledge can guide vision models to handle specific tasks. Classical computer vision models focus on tasks in natural image scenes and are not good at those in heterogeneous image scenes. Heterogeneous image knowledge created and exploited from the perspective of focusing on heterogeneity enables visual perception models to generalize well from natural image scenes to heterogeneous image scenes.

Visual perception framework based on heterogeneous image knowledge: Considering the specificity of visual perception tasks oriented to heterogeneous images, we propose a visual perception framework based on heterogeneous image knowledge. As shown in Fig. 3, the proposed framework is driven by both representation and knowledge. On the one hand, visual representation is constructed through representation extraction and representational learning based on heterogeneous visual data; on the other hand, knowable representation is explored and established for visual tasks related to specific

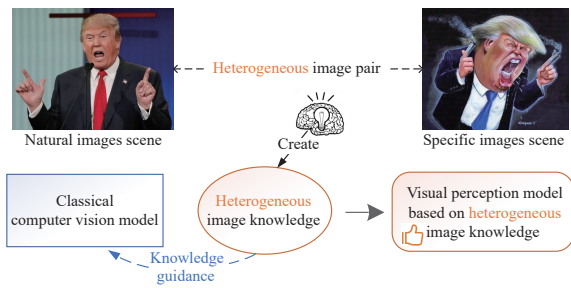


Fig. 2. The key idea of our framework.

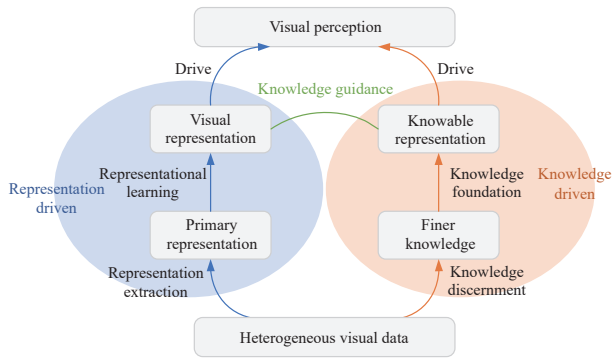


Fig. 3. The proposed framework for visual perception based on heterogeneous image knowledge. Different from traditional frameworks [4], our framework is driven by knowledge in addition to representation. Exploring and building knowable representation for visual tasks related to specific heterogeneous images, and making the knowledge accessible to visual models through the knowledge guidance mechanism can lead to better solutions of heterogeneous-related visual perception problems.

heterogeneous images. Then, visual representation and knowable representation are combined to effectively solve the visual problems related to heterogeneous images. Our framework aims to obtain task-specific heterogeneous image knowledge through human experience summarization or knowledge mining, and establish a knowledge-guided mechanism to fuse these knowledge into a generic visual perception model to deepen the understanding of a specific task. Thus, the visual problems related to heterogeneous images can be addressed more effectively. The generic visual perception models mentioned here refer to the common visual models currently used in various computer vision tasks.

The part of representation driven: This part aims to obtain effective visual representations to support visual perception tasks for heterogeneous images. Particularly, for heterogeneous images in a specific vision task, a coarse primary representation is first obtained by representation extraction. The primary representation is then further processed using representational learning to obtain a visual representation that satisfies the demands of a concrete task. There are two kinds of methods for obtaining visual representations, including explicit and implicit methods.

Explicit methods usually use hand-designed feature extractors to obtain explicit image features first, and then adopt principal component analysis (PCA), linear discriminant analysis (LDA) and other representational learning techniques to acquire the desired visual representations. Hand-designed image features, i.e., handcrafted features, are diverse and include both global features such as color histograms and local features such as corner detection operators.

Although the explicit method is completely transparent to the designer and easy to make targeted adjustments and improvements when visual representations are acquired explicitly, it is entirely manual. To obtain a good visual representation requires not only extensive expertise, but also a lot of time and effort for the designer to repeatedly test and adjust. While implicit approaches represented by deep neural networks can learn the visual representation required for a task directly from the image. Thanks to the powerful representational and modeling capabilities of deep neural networks, they are able to automatically learn visual representations that meet the

requirements of specific tasks. Therefore, implicit automatic representation learning techniques based on deep learning have gradually replaced the explicit visual representation extraction process defined by humans. Nonetheless, in some scenarios where the computational capacity is very limited or the real-time requirement is high and the performance requirement is relatively low, the manually defined visual representation extraction process still has its applications. Therefore, a comprehensive consideration is needed for the specific approach to obtain visual representations when dealing with practical vision problems.

The part of knowledge driven: In this part, the goal of knowledge discernment is to explore and refine the knowledge associated with a specific heterogeneous image task. While knowledge foundation is intended to build the discovered heterogeneous image knowledge into the knowable representation that can be applied by computer. This process can be realized concretely by means of heterogeneous image knowledge model. In general, human knowledge can be divided into explicit knowledge and tacit knowledge [5]. Explicit knowledge can be expressed explicitly through written words, formulas, diagrams, etc., and can be easily taught, such as laws of physics, mathematical logic, and artificially defined rules. Tacit knowledge is often implicit in data and may change due to alteration in the scale of data, which needs to be acquired through training and can not be taught directly, such as the rise and fall of oil prices. The knowledge can be either explicit or implicit after knowledge discernment.

Knowledge foundation is not only the process of representing the discovered knowledge relevant to a particular heterogeneous image task in a computer-understandable way, but is also about building it into knowable representation that can be combined with visual features. Representation extraction and representational learning map images from the data space to the feature space. Therefore, knowledge foundation also involves transforming the uncovered knowledge related to a specific heterogeneous image task into feature space for subsequent knowledge guidance.

Heterogeneous image knowledge driven visual perception: After getting both visual and knowable representations, knowledge guidance enables the knowledge to be accessed by the visual model, so as to obtain the visual perception model for specific heterogeneous images with better performance. Knowledge guidance is able to combine knowable representation and visual representation, and then through constraints and joint training, they can achieve some coordination and unity to jointly drive and promote visual perception for heterogeneous images.

Knowledge guidance can be achieved through various concrete methods. For example, computer-available heterogeneous image knowledge representations can be introduced into a visual network, and the objective function of the representations can be optimized by iterative training to close the intra-class distance and increase the inter-class difference. In addition, loss functions can be designed for the similarity of knowable representation and visual representation results for a specific heterogeneous image visual perception task to induce consistent similarity between the two representations. Moreover, the unified constraint on knowledge and visual information can be realized by using the supervision information of the whole visual perception model oriented to specific heterogeneous images.

Case study: We conduct case studies to show how the proposed framework can be used for visual perception tasks related to heterogeneous images. Specifically, two tasks, single image dehazing and face photo-sketch synthesis, are taken as examples, and two models are designed to solve these two problems from the characteristics of each problem, combined with heterogeneous image knowledge.

Solving the problem of single image dehazing: As an ill-posed problem, single image dehazing is exceedingly challenging. In this task, the haze image and the clear haze-free image are a set of heterogeneous image pairs, and there exists some heterogeneous image knowledge that can be explored. In particular, it is not difficult to observe that the distribution of haze in the images is often uneven. Additionally, it has been demonstrated that dark channels exist in non-sky local areas of the image. Based on these perspectives, we propose a dehazing model based on heterogeneous image knowledge, i.e., a feature aggregation attention network (FAAN) [6], which can adaptively aggregate features at different levels and recover a clear version of a given haze image. Fig. 4 shows an example of the appli-

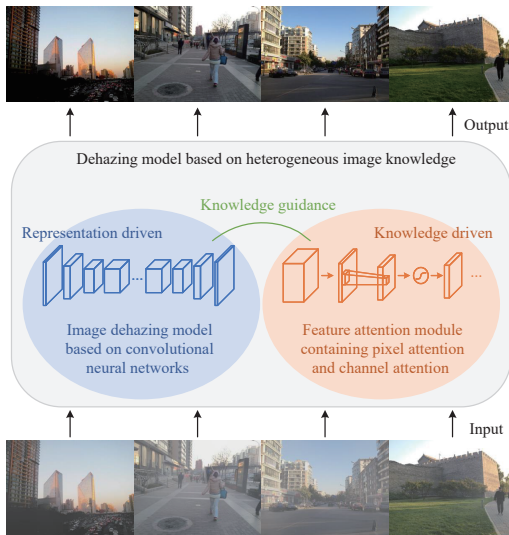


Fig. 4. The application example of our framework on single image dehazing.

ation of the proposed framework on the image dehazing task.

As displayed in Fig. 4, on the one hand, we introduce an image dehazing model based on convolutional neural networks to acquire visual representations; on the other hand, a feature attention module containing pixel attention and channel attention is designed to obtain knowable representations. The pixel attention and channel attention modules make the model pay more attention to thick haze pixels and feature channels that are of more importance for the dehazing task. Visual representation and knowledge representation are integrated through knowledge guidance mechanism, and the two are driven together to solve the image dehazing problem.

Solving the problem of face photo-sketch synthesis: Face photo-sketch synthesis consists of two tasks: generating face sketches from real photos and transforming sketch images into face photos. In this task, the face photo and face sketch constitute a group of heterogeneous image pairs. As one of the major research objects in biometric recognition, faces naturally imply a wealth of identity information. Therefore, we propose a face photo-sketch synthesis model based on heterogeneous image knowledge, namely, an identity-sensitive generative adversarial network (IsGAN) [7], which is able to synthesize face sketches and face photos that retain identity details. Fig. 5 gives an example of the heterogeneous image knowledge based visual perception framework applied to face photo-sketch synthesis tasks.

As can be seen in Fig. 5, for one thing, a face photo-sketch synthesis model based on generative adversarial networks is designed to capture visual representations; for another, we introduce additional identity labels and propose an identity recognition loss function to obtain knowledge representations. Combined with the knowledge guidance mechanism, visual representations can be interwoven with knowledge representations, and the face sketch synthesis task is finally accomplished through experiments and evaluations. In addition, numerous experiments demonstrate that the introduced identity recognition loss function contributes to the retention of identity information in the synthesis process, thus improving the quality of the generated results.

Summary: The proposed framework expects that the introduction of heterogeneous image knowledge can improve existing classical computer vision models to solve heterogeneous image oriented visual perception problems. The case study performed on this basis shows that the proposed framework is successful in practice for single image dehazing and face photo-sketch synthesis tasks, and can provide a reference for designing other visual perception models related to heterogeneous image. The results of our designed dehazing and face photo-sketch synthesis models also indicate that the introduced heterogeneous image knowledge is effective and capable of improving the performance of the model. Therefore, combining heterogeneous image knowledge with classical visual perception models that

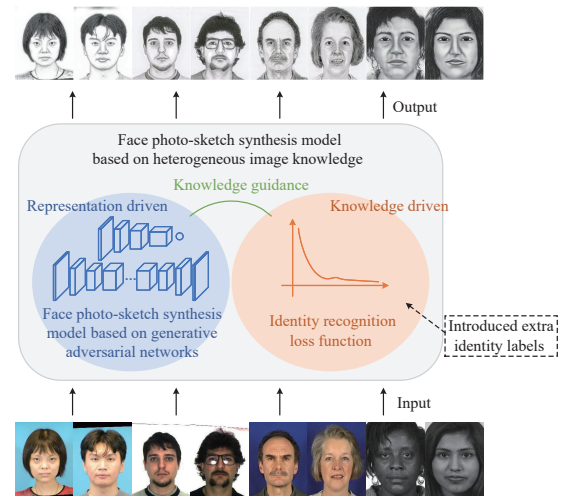


Fig. 5. The application example of the proposed framework on face photo-sketch synthesis. Here the input is a photo and the output is a sketch, the model can also synthesize a photo given a sketch.

are not adapted to heterogeneous image related tasks can deepen the understanding of specific tasks and thus enable visual perception of heterogeneous images.

Conclusion: The emergence of a large number of heterogeneous images brings new challenges to visual perception models. Due to the great differences between different image modalities, the traditional computational vision models focusing on natural scene images are difficult to be used to tackle the visual perception problems of heterogeneous images. In this context, this letter proposes a novel framework for visual perception based on heterogeneous images knowledge. The purpose of this framework is to explore heterogeneous image knowledge for specific tasks and integrate it into classical visual perception models to enhance the understanding of specific tasks, so as to address visual perception problems related to heterogeneous images more effectively. In addition, the models designed based on the proposed framework are explored and studied on the tasks related to heterogeneous images, such as image dehazing and face photo-sketch synthesis, and the results demonstrate that the framework can achieve excellent performance on both tasks.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (62302161, 62303361), and the Postdoctoral Innovative Talent Support Program of China (BX20230114).

References

- [1] X. Hong, T. Zhang, Z. Cui, and J. Yang, "Variational gridded graph convolution network for node classification," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 10, pp. 1697–1708, 2021.
- [2] Y. Qiu, Z. Lu, and S. Fang, "A short-term precipitation prediction model based on spatiotemporal convolution network and ensemble empirical mode decomposition," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 4, pp. 738–740, 2022.
- [3] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Computational knowledge vision: Paradigmatic knowledge based prescriptive learning and reasoning for perception and vision," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 5917–5952, 2022.
- [4] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. MIT Press, 1989.
- [5] H. Collins, *Tacit and Explicit Knowledge*. Chicago, USA: University of Chicago Press, 2010.
- [6] L. Yan, W. Zheng, C. Gou, and F.-Y. Wang, "Feature aggregation attention network for single image dehazing," in *Proc. IEEE Int. Conf. Image Processing*, 2020, pp. 923–927.
- [7] L. Yan, W. Zheng, C. Gou, and F.-Y. Wang, "IsGAN: Identity-sensitive generative adversarial network for face photo-sketch synthesis," *Pattern Recognition*, vol. 119, p. 108077, 2021.