Xin Cai xincai00@gmail.com

Research Proposal

Xin Cai

Title: Adversarially Learned Robust Deep Representations for Generalizable Person Re-identification

1. Introduction

Person re-identification (Re-ID) has been an active research area in computer vision, aiming at retrieving the person of interest across multiple non-overlapping cameras given a query image or video clip. It has great potential in intelligent surveillance applications and human-robot interactions. Re-ID is a challenging task due to the significant intra-domain variation, such as different viewpoints, occlusions, unconstrained poses, and background clutter, and the severe performance degradation when deployed to domains distinctive from the training setting. Generally, Re-ID systems can be subsumed under two broad categories: closed-world and open-world settings. The former refers to testing scenarios are identical or highly similar to those encountered during the process of training (e.g., data are collected from the same camera networks), and the latter is also called cross-domain Re-ID, meaning that there is a huge domain gap between training and testing. With advances made in deep learning, recently proposed Re-ID systems have achieved promising results, especially on the closed-world setting where performance has virtually become saturated.

However, the idealistic closed-world setting postulates that training and testing data are drawn from the same camera network or the same domain, which rarely holds in the real-world deployment, therefore severely restricting the applicability of such domain-specific Re-ID models to real-world scenarios. Therefore, many research efforts have to be made in order to narrow the gap between the closed-world and open-world applications, taking a step towards real-world Re-ID system design. This will also be the primary focus of my research, i.e., developing Re-ID systems that generalize well to unseen scenarios (cross-domain Re-ID). Specifically, in the cross-domain Re-ID setting, source and target domains contain disjoint identity classes and dramatic differences caused by the domain gap, such as background, viewpoint, and illumination. Besides, identity labels are unavailable in the target domain due to the prohibitively high cost of manual annotation. As a result, learning robust feature representations is at the core of developing generalizable Re-ID systems. To be specific, the remarkably high accuracy of current Re-ID systems developed in the closed-world setting comes at the cost of fitting undesired attributes of training data, i.e., the well-known overfitting problem. Consequently, robust feature learning requires effectively regularize the feature learning process so that it can focus on identity-related characteristics rather than other identity-unrelated cues which will present significant variance across different domains and inevitably introduce interference, i.e., distilling domain-invariant

feature representations. One of the most promising research direction that can realize the purpose mentioned previously is adversarial learning. Concretely, existing approaches using adversarial learning to improve the generalization ability of Re-ID systems generally follow two different principles. One is characterized by using generative models to perform data augmentation, and the other focuses on designing various adversarial losses to regularize the feature learning process, e.g., by disentangling the identity-relevant components from identity-irrelevant components.

In addition to the above-mentioned challenges, conventional Re-ID systems are mainly image-based. Extending these image-based Re-ID systems to successfully process video data needs to tackle obstacles caused by temporal variations. More specifically, the temporal motion would bring intra-class variance like the change of poses, especially causing difficulty for temporal feature aggregation. Last but not least, the real-world deployment of Re-ID systems would benefit from advances in few-shot learning or meta-learning. For example, the capability of trained Re-ID systems being directly applied to unseen scenarios without model updating or very few numbers of model updates is crucial due to the limited computational budget on mobile or edge devices. However, current methods addressing cross-domain Re-ID require extensive model parameter updates jointly using labelled source domain and unlabelled target domain data samples, which is computationally prohibitive for mobile applications. Therefore, leveraging advances in few-shot learning or meta-learning to develop generalizable Re-ID systems applicable to computational resource-restricted platforms is of significant practical value.

To summarize, the primary research objectives are as follows:

- developing methods to regularize the feature extracting process to obtain disentangled feature representations such that identity-irrelevant factors could be considerably removed.
- developing methods specifically considering temporal variations to distil temporal coherent feature representations for video-based Re-ID systems.
- developing generalizable Re-ID systems by leveraging advances in few-shot learning or meta-learning to achieve computationally efficient and rapid deployment for unseen scenarios.

2. Related Work

2.1 Conventional Deep Re-ID Methods

Conventional deep Re-ID models refer to those proposed in the closed-world setting. A large family of these person Re-ID models focuses on supervised learning. They usually approach Re-ID as deep metric learning problems[1], exploit pedestrian attributes as extra supervision signal via multitask learning[2, 3], utilize part-based matching or ensembling to reduce intra-class variations[4, 5], or make use of human pose and parsing to facilitate local feature learning[6, 7].

2.2 Unsupervised Domain Adaptation

Despite the significant performance improvement obtained in the closed-world setting, it has been reported that deep Re-ID models trained on the source domain are likely to encounter dramatic performance degradation when deployed on a novel target domain. This largely restricts the applicability of such domain-specific Re-ID models, consequently having received increasing research attention in recent years. Current cross-domain Re-ID systems are mostly motivated by approaches proposed in unsupervised domain adaptation (UDA), which employs labelled data in the source domain and unlabelled data in the target domain to improve the model performance in the target domain. The fundamental principle of most UDA models is to align feature distributions between source and target domains, e.g., by adversarial training or by minimizing certain types of distance or divergence measures. Correspondingly, UDA methods can be roughly categorized as input-level or feature-level adaptation. At the input-level, models are usually adapted by training with style translated images[8, 9, 10]. Adaptation at feature-level often minimizes certain distance or divergence measures between source and target feature distributions, such as correlation[11], maximum mean discrepancy (MMD)[12, 13]. CyCADA[14] adapts at both input-level and feature-level to combine benefits from both worlds. Besides, several methods utilize the similarity of features to generate pseudo-labels for unlabelled target samples[15, 16]. In[16], an approach is presented to estimate the labels of unlabelled samples by using the k-nearest neighbours. Then, the predicted labels are further used to learn optimal deep representations.

2.3 Cross-Domain Person Re-ID

Most of existing unsupervised domain adaptation methods assume that class labels remain the same across domains, while the person identities (classes) of different Re-ID datasets are entirely different (disjoint). Hence, the approaches mentioned above fail to be directly used to solve the problem of unsupervised domain adaptation in person Re-ID. Current cross-domain person Re-ID models can be generally classified into three broad categories: 1) translation-then-learning paradigm; 2) using adversarial training to distil disentangled feature representations; 3) self-training by using generated pseudo labels for target domain samples. Models following the translation-then-learning paradigm[17, 18, 19, 20] consist mainly of two steps: 1) using conditional generative adversarial networks (GANs) to translate the annotated source dataset to target domains in an unsupervised fashion meanwhile preserving identity-related information as much as possible; 2) then using the translated dataset with reduced domain gap to perform supervised learning. Besides, most models realize domain adaptation based on frameworks proposed in unpaired image-to-image translation, such as CycleGAN[21] and StarGAN[22]. While another line of research tackling cross-domain person Re-ID is also reliant on adversarial learning, it is characterized by leveraging various adversarial losses to regularize feature extracting functions such that identity-irrelevant information can be considerably stripped off. The representative work includes FD-GAN[23], DGNet[24], DGNet++[25], etc. The translating-

then-learning paradigm can be regarded as performing data augmentation based on deep generative models, but it is a process separated from Re-ID feature learning. Compared to them, integrating adversarial learning into the Re-ID feature representation extraction process is more beneficial, leading to disentangled and hence robust feature representations. Apart from methods based on generative models, there are many methods addressing cross-domain person Re-ID based on the self-training principle[26, 27]. It produces pseudo labels for target domain samples based on clustering methods, and then performs clustering and supervised learning alternately to progressively enhance the model performance on the target domain.

2.4 Video-based Person Re-ID

Video sequences provide more abundant and diverse identity-related information, meanwhile introducing temporal interference. For learning more robust representations from video sequences, existing video-based person Re-ID methods mainly take great efforts to 1) mine the motion clues in the person video; 2) aggregate the video sequence embeddings. To extract discriminative motion clues, many methods[28, 29, 30] model the motion process with the recurrent neural networks (RNNs) or long short-term memory networks (LSTMs). Recently, 3D convolution neural networks (3D-CNNs)[31, 32] have been applied for video person Re-ID to jointly learn the appearance and motion clues. While the idea of decomposing content from motion for video generation has been explored in the context of deep generative modelling [33], it has yet not been systematically studied to mitigate the negative effect caused by temporal variance in video-based person Re-ID. The pioneering work along this direction is proposed in [34], where authors attempted to explicitly decompose video feature representations into temporal coherent and motion components, respectively.

2.5 Few-shot Learning

Despite that fact that UDA-based cross-domain Re-ID models have achieved promising results, the extensive parameter updates of UDA-based models require using unlabelled target domain data samples, meaning that data collection and model update are indispensable. The domain generalization problem has also been investigated in the few-shot learning setting, aiming at grasping novel concepts based on very few data samples. The research progress of few-shot learning has been slow due to its intrinsic difficulty. However, the resurgence of meta-learning in deep learning era has sparked considerable interest in recent years. Roughly speaking, the key idea is that meta-learning agents improve their own learning ability over time, or equivalently, learn to learn. The learning process is primarily concerned with tasks (set of observations) and takes place at two different levels: an inner- and an outer-level. At the inner-level, a new task is presented, and the agent tries to quickly learn the associated concepts from the training observations. This quick adaptation is facilitated by knowledge or experience that has been accumulated across earlier tasks at the outer-level. There have been initial attempts[35] of formulating cross-domain Re-ID within the framework of few-shot learning, therefore benefiting from advances in this field, such

as Matching Network[36], Prototypical Network[37], and graph neural network-based few-shot learning[38].

3. Methodology

The literature review section has covered representative solutions used to improve the generalization ability of Re-ID systems. The research will be conducted by revolving around three primary objectives put forward in the introduction section and following current mainstream solutions.

3.1 Adversarially Learned Robust Representations for Image-based Re-ID

As stated in the literature review section, adversarial learning-based methods can be generally categorized as 1) the translation-then-learning paradigm; 2) disentangled feature representation distilling. There are several improvements that can be further made. Firstly, current methods are dominated by GANs. With rapid progress in deep generative models, there are many appealing alternatives worthy to be explored to verify their effectiveness in cross-domain Re-ID, such as variational autoencoders (VAEs)[39], deep autoregressive models[40], flow-based models[41], and hybrid models[42, 43, 44]. These methods have demonstrated their effectiveness in content generation and disentangled feature learning. For example, adversarial autoencoder (AAE)[43] uses GAN to perform variational inference such that arbitrary prior distributions can be imposed, therefore enabling its broad applicability in semi-supervised learning, disentangling style and content of images, and unsupervised clustering. Many research efforts have been devoted to learning more interpretable factorized features in an unsupervised manner, such as InfoGAN[45] and β-VAE[46]. All these advances can be further explored for either increasing the controllability of image generation process or disentangled feature learning, leading to more generalizable Re-ID models. Besides, advances in person image synthesis, especially pose-guided person image synthesis[47, 48], can provide inspirations for adversarial learning-based data augmentation or feature extraction. For example, DGNet and DGNet++ decompose feature representations into two subspaces: appearance and structure subspaces. While achieving satisfactory disentanglement to some extent, the appearance encoder only focuses on extracting information related to clothing and shoes. Therefore, realizing feature disentangling at a more fine-grained level is a promising research direction, e.g., distilling pose-independent representations. Different from general image translation tasks in which object structures or spatial layouts are relatively unchanged, person image synthesis poses great challenges due to pixel-to-pixel misalignment or structural deformation between conditioning and target images. There have been several attempts[23, 49] using pose-guided person image synthesis to enhance the robustness of Re-ID systems against pose variations. Last but not least, using adversarial learning to improve the generalization ability of Re-ID systems is an emerging research direction. Although it has drawn increasing research attention

and shown promising results, current methods still lack systematic investigation. For example, a variety of adversarial losses has been proposed, but these losses have not been rigorously evaluated in ablation experiments, causing difficulty for fair comparisons. How to establish a simple but strong baseline model is of significant importance.

3.2 Adversarially Learned Robust Representations for Video-based Re-ID

With the rapid progress in image-based person Re-ID, video-based Re-ID has gradually drawn increasing research attention as a natural extension. Apart from dealing with difficulties existing in image-based Re-ID, video Re-ID needs to tackle challenges resulted from temporal variations, such as occlusions, visual appearance changes, pose changes and ambiguities caused by motion. Various temporal variation factors may cause inconsistency over the temporal dimension and corrupt discriminative features extracted from individual video frames, therefore posing great challenges for learning robust representations for video sequences. Existing methods focus mainly on designing robust temporal feature aggregation schemes that can mitigate the adverse effect caused by temporal variations, such as spatial-temporal attention models[50]. Currently, decomposing temporal variation factors from identity-related clues has seldom been studied. One of the pioneering research work along this direction is [34], where authors have attempted to use adversarial learning to explicitly decompose video representations into temporal coherent and motion parts respectively based on the observation that temporal motion features may introduce extra noise. Apparently, extracting feature representations invariant to temporal variations using adversarial learning is similar in spirit to its static counterpart, which aims to enhance feature robustness against variations caused by viewpoints, poses, and camera configurations. As a result, how to extend successful adversarial learning strategies employed in image-based Re-ID to video domain and combine advances in video generation[33, 51] or spatial-temporal predictive learning[52] is a potentially reasonable research direction to extract robust representations for video-based Re-ID.

3.3 Few-shot Re-ID

While UDA-based models have demonstrated their effectiveness in adapting Re-ID models from source to target domains, some researchers have pointed out that UDA-based models still require data collection and model updates[35], which may limit their applicability in the real-world deployment. Consequently, they attempt to bridge the gap using methods proposed in few-shot learning or meta-learning. One of the primary objectives of few-shot learning is to endow machines with human-like intelligence–the ability to quickly grasp a novel concept with a single or handful of examples. Recently, deep meta-learning techniques have been actively studied and shown promising results, which can be roughly subsumed under three categories: 1) metric-[36, 37, 38], 2) model-[53, 54], 3) optimization-based[55, 56] meta-learning techniques. It is beneficial to regard Re-ID models trained using UDA techniques as pre-trained models, and

then leveraging advances in deep meta-learning to further improve the generalization ability.

4. Timeline

Generally, I plan to divide the period of three years of working towards a PhD degree into two parts: 1) the first two years will focus on publishing three papers on international conferences or journals with potential research topics surrounding around those stated in section 3.1, 3.2, and 3.3, respectively; 2) the final year will be used to complete my PhD dissertation by summarizing the previous work.

The detailed timetable of the first two years are outlined as follows:

- $1 \sim 4$ months: conducting a comprehensive literature review and selectively reproducing results of some representative research work, revolving around the central topic: image-based cross-domain Re-ID systems using adversarial learning techniques;
- $5 \sim 8$ months: proposing my own improvements and publishing on international conferences or journals, striving to establish a simple but strong baseline model by conducting rigorous ablation studies;
- $9 \sim 12$ months: conducting a comprehensive literature review and selectively reproducing results of some representative research work, revolving around the central topic: video-based cross-domain Re-ID systems using adversarial learning techniques;
- $13 \sim 16$ months: proposing my own improvements and publishing on international conferences or journals with a particular focus on distilling feature representations invariant to temporal variations;
- $17 \sim 20$ months: conducting a comprehensive literature review and selectively reproducing results of some representative research work, revolving around the central topic: few-shot or meta-learning-based Re-ID systems;
- $21 \sim 24$ months: proposing my own improvements and publishing on international conferences or journals;

References

[1] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[2] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 475–491.

[3] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 365–381.

[4] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 420–428.

[5] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1077–1085.

[6] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[7] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.

[8] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3722–3731.

[9] A. Dundar, M.-Y. Liu, T.-C. Wang, J. Zedlewski, and J. Kautz, "Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation," *arXiv preprint arXiv:1807.09384*, 2018.

[10] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.

[11] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[12] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[13] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2272–2281.

[14] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[15] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Advances in neural information processing systems*, 2013, pp. 46–54.

[16] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," *Advances in Neural Information Processing Systems*, vol. 29, pp. 2110–2118, 2016.

[17] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[18] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 994–1003.

[19] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–188.

[20] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7202–7211.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[22] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[23] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in neural information processing systems*, 2018, pp. 1222–1233.

[24] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 2138–2147.

[25] Y. Zou, X. Yang, Z. Yu, B. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," *arXiv preprint arXiv:2007.10315*, 2020.

[26] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6112–6121.

[27] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8222–8231.

[28] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*. Springer, 2016, pp. 701–716.

[29] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, "Video-based person re-identification with accumulative motion context," *IEEE transactions on circuits and systems for video technology*, vol. 28, no. 10, pp. 2788–2802, 2017.

[30] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4747–4756.

[31] J. Li, S. Zhang, and T. Huang, "Multi-scale 3d convolution network for video based person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8618–8625.

[32] X. Liao, L. He, Z. Yang, and C. Zhang, "Video-based person re-identification via 3d convolutional networks and non-local attention," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 620–634.

[33] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.

[34] G. Chen, Y. Rao, J. Lu, and J. Zhou, "Temporal coherence or temporal motion: Which is more critical for video-based person re-identification?" in *European Conference on Computer Vision*. Springer, 2020, pp. 660–676.

[35] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 719–728.

[36] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[37] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[38] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[39] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[40] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[41] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[42] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *arXiv preprint arXiv:1505.05770*, 2015.

[43] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[44] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," *arXiv preprint arXiv:1706.04987*, 2017.

[45] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, pp. 2172–2180, 2016.

[46] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[47] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356.

[48] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 717–734.

[49] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4099–4108.

[50] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 369–378.

[51] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu, "Free-form video inpainting with 3d gated convolution and temporal patchgan," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9066–9075.

[52] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International Conference on Learning Representations*, 2018.

[53] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, 2016, pp. 1842–1850.

[54] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," *arXiv preprint arXiv:1707.03141*, 2017.

[55] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.

[56] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *arXiv preprint arXiv:1703.03400*, 2017.