How do you Perceive Differently from an AI — A Database for Semantic Distortion Measurement

Shuxin Zhao*, Jiahua Xu*, Yongquan Hu, Wei Zhou, Sen Liu, Zhibo Chen†
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China, Hefei 230027, China
Email:{sx2222,xujiahua,yongquanh,weichou}@mail.ustc.edu.cn, elsen@iat.ustc.edu.cn, chenzhibo@ustc.edu.cn

***Abstract*—Artificial intelligence (AI) is enabling the automated analysis of large amounts of image/video data, boosting the speed of multimedia data processing remarkably. Meanwhile, Image Quality Assessment (IQA) plays an important role in developing automatic analysis methods. To ensure the effectiveness of AI, images in multimedia applications should be considered for visual examination by both human and machine. Therefore, it is significant to understand the differences between human's and AI's perception of semantic distortion. However, little work has been done due to the lack of data from human on the semantic level. In this paper, we first propose a semantic database (SID) based on the surveillance scenarios, by collecting subjective average recognition rates of 3 semantic targets (face, pedestrian, license plate) with 3 types of distortion (JPEG Compression, BPG Compression, Motion Blur). Then, we present a detailed analysis of how human and AI perceive semantic distortion differently. Experimental results show that AI is stronger in tolerance to distortion than human beings on average, while weaker at generalization and stability. It is also implied in the experiments that existing IQA methods are not effective enough at judging the semantic distortion.**

***Index Terms*—image quality assessment, image semantic analysis, image recognition, surveillance scenario**

I. Introduction

Artificial intelligence (AI) is rapidly developing as substitutes for human in areas such as classification, recognition and categorization. Such high-level semantic tasks enable automatic analysis of large amounts of image/video data, which extraordinarily enhances efficiency and effectiveness. While developing these analysis methods, Image Quality Assessment (IQA) metrics that simulate human judgments are widely used for evaluation or guidance. However, inconsistency between human judgments and AI can influence the performance of these methods, such as haze removal [1]. Therefore, one significant problem is to understand how human and AI perceive semantic distortion differently.

Previous subjective quality datasets like LIVE, TID2013, CSIQ, BAPPS generally focus on low-level quality judgments such as Mean Opinion Score (MOS) [2]–[4], Two Alternative Forced Choice (2AFC) or Just Noticeable Differences (JND) [5]. However, low correlations between low-level perceptual tasks and high-level semantic tasks have been demonstrated on the BAPPS database [5]. Few databases are available yet to provide data on human perception in high-level recognition tasks, which is essential to the problem.

In this paper, we establish the **S**eman**t**ic **D**atabase (SID) to lay the foundation for investigating the perceptual differences between human and AI. Since human judgments are highly context-dependent [5], we build the database based on the surveillance scenarios so we can narrow down the distribution of semantic information to specific targets, and link semantic distortion to the accuracy of the corresponding high-level recognition task. Surveillance cameras deployed in public places like airports, college campuses or office buildings provide huge amounts of video data for identity authentication or suspicious activity detection. Since the position of the camera is fixed and the background remains unchanged, it is generally assumed that the majority of visual semantic information lies in the high-value targets in the foreground.

Face, license plate, pedestrian are the most important targets, so we include a subset for each of them in the database. As one of the most reliable and accessible biometric features, face is significant for identity authentication concerning military, finance, and public security. Previous works [6]–[13] have boosted accuracy to above 99% on the most popular benchmark Labeled Faces in the Wild, surpassing average human accuracy of 97.53% [14]. License Plate Detection [15]–[19] and Recognition [20]–[24] have various applications in public security, traffic safety and commercial scenarios. Person re-identification [25]–[28] is another fundamental task in the automated video surveillance system, which is aimed at establishing consistent labeling across multiple cameras to recover disconnected or lost tracks.

We then apply three types of distortions in the database: Joint Photographic Experts Group (JPEG) compression, Better Portable Graphics (BPG) compression and motion blur. During video shooting, transmission or storage, semantic information in the video data are prone to degradation due to limitations of resolution or exposure time, transmission errors and compression. JPEG compression is the most widely-used compression format, and BPG compression is included as a representative of the High Efficiency Video Coding (HEVC) standard. Motion Blur occurs when objects in the image changes during the recording of a single exposure due to rapid movement, such as vehicles driven at high speed.

Based on the proposed database, we conduct experiments to seek insight into differences in the perception of semantic

*Equal Contribution

†Corresponding author: Zhibo Chen (chenzhibo@ustc.edu.cn). This work was supported in part by NSFC under Grant 61571413, 61632001,61390514.

distortion between human and AI. It is demonstrated in the experiments that models have a remarkable advantage over human when high distortions are involved. However, models are unstable when distortion is alleviated, which means they could produce correct results for severe distortions but sometimes not for the less severe. We also show that IQA metrics could not reliably and efficiently measure the influence of distortions for the aforementioned high-level recognition tasks.

In Section II, we describe the details of building the Semantic Database. In Section III, we compare the perception of semantic distortion between human and state-of-art models and evaluate the performance of existing IQA methods. In Section IV, we summarize the paper and inspirations for future work.

II. Semantic Database

This section introduces the subjective experiment methods of the semantic database. In order to measure the perceptual differences between human and AI, we build SID which includes 100 reference face, pedestrian and license plate images, respectively. The database will be published for future research on semantic distortion.

A. Selection of Images

The semantic database SID contains three subsets: face, pedestrian and license plate image set, each with 100 reference images. Reference images for the face are collected from LFW [14], and from Market-1501 [29] for the pedestrian. Two images from the same identity are respectively used as the test and template image as shown in Fig.1. Differently, license plate images are selected from the Internet and the database is built by ourselves.

(a) (b) (c) (d)

Fig. 1: Examples of test and template images in the face and pedestrian subset. (a) Face template image. (b) Face test image. (c) Pedestrian template image. (d) Pedestrian test image.

B. Distortion Processing

Recognition of the image content by human subjects is disturbed when distortions are introduced. To imitate this situation, three kinds of distortion are involved in the database, including JPEG compression, BPG compression and motion blur. JPEG is the most common compression method and BPG can be regarded as a representative of the HEVC standard. Both of them are chosen to simulate the artifacts induced by compression. Motion blur is generated while there exists relative motion between the camera and the object during shooting. Since objects in the scene are usually moving when photographing license plate or pedestrian images, motion blur is introduced to simulate motion in real scenarios.

To investigate the relationship between distortion level and recognition accuracy, the distortion parameters of 20 distortion levels are carefully chosen as listed in Table I. The parameters are capable of covering a quality range that corresponds to a wide span of recognition accuracy. Image quality touches the bottom when QP value equals 51 or quality parameter equals 1. However, there is no boundary for the maximum kernel size of motion blur. Preliminary tests are conducted first to decide the boundary of the kernel size at which subjects can hardly recognize the content correctly.

C. Subjective Test Methods

During the subjective experiment, the subject is asked to look at the distorted images and choose if he can recognize. As shown in Fig.2, the answers can be divided into three categories: cannot recognize, can recognize but wrong, can recognize and right. Specifically, each subject can only see one specific distortion type for each reference image since human beings have memories. Otherwise, the subject can recognize the content in the image even if it is blurry once he has seen the clear version before, which violates the principle of the experiment.

For the face subset, images pairs are manually divided into 10 groups by appearance similarity: *sports women, long brown hair woman, long blonde hair women, short brown hair women, short blonde hair women, sports men, black men, bald man, brown hair men*, and *blonde hair men*. Such grouping is performed to enforce recognition from facial features rather than gender, hairstyle or complexion. Similarly, the pedestrian subset is divided into eight groups according to colors. As for the license plate subset, subjects need to recognize the entire license plate number sequence, including provincial abbreviation, numbers and letters.

TABLE I: Distortion parameters of 20 distortion levels for different distortion types (Ref means the reference image)

Distortion level		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
BPG (*QP*)		51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	30	20	10	Ref
JPEG (*Quality*)		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	40	60	80	Ref
Motion blur (*Kernel Size*)	Face	50	45	40	35	30	28	26	24	22	20	18	16	14	12	10	8	6	4	2	Ref
	Pedestrian	30	20	18	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	Ref
	License	20	16	12	10	9	8	7.5	7	6.5	6	5.5	5	4.5	4	3.5	3	2.5	2	1	Ref

Fig. 2: Graphical interface of our experiment for face recognition.

62 non-expert subjects aged from 21 to 49 take part in our experiment. All participants pass the visual acuity and color vision (Ishihara charts). In the formal test, every subject is presented with three subsets including 300 image groups in a random order. Each image group includes distorted versions of the same reference image at 20 levels.

III. Analysis of Subjective Database

This section provides a detailed analysis of the subjective results. First, outlier removal is performed to guarantee the effectiveness of data collected from subjects. Second, the variations and consistency in recognition ability among subjects are analyzed. Third, we explore the differences between human beings and deep learning models. Finally, evaluation of some objective image quality assessment metrics is performed on the semantic database to see whether they are suitable for measuring semantic distortion.

A. Outlier Removal

Before the subjective experiment, there were originally 110 reference images in each subset to leave a margin for weird samples. Images too hard to recognize even clear or too easy to recognize even distorted are removed to ensure the usability of the semantic database, finally remaining 100 reference images in each subset. Then, subjects with comparatively low accuracy for reference images are also excluded. 60 subjects remain valid after removal.

B. Variations and Consistency among Subjects

Recognition accuracy of an image is computed by averaging results of valid subjects. The increase of distortion is expected to have a loss of accuracy by the human being. According to Fig.3, the recognition accuracy only starts to drop dramatically when QP is greater than 35 and when quality is smaller than 15. Contrastly, for motion blur the decline of the accuracy is much smoother but starts at a slight degree.

Average of the accuracy at each distortion level reflects the consistency of subjects when identifying the images, while maximum and minimum of the accuracy indicate the discrepancy among them. To better illustrate the differences, we draw the histogram for each subset and distortion type. First, for each image, the distortion level at which the subject begins to recognize correctly is recorded and addressed as the

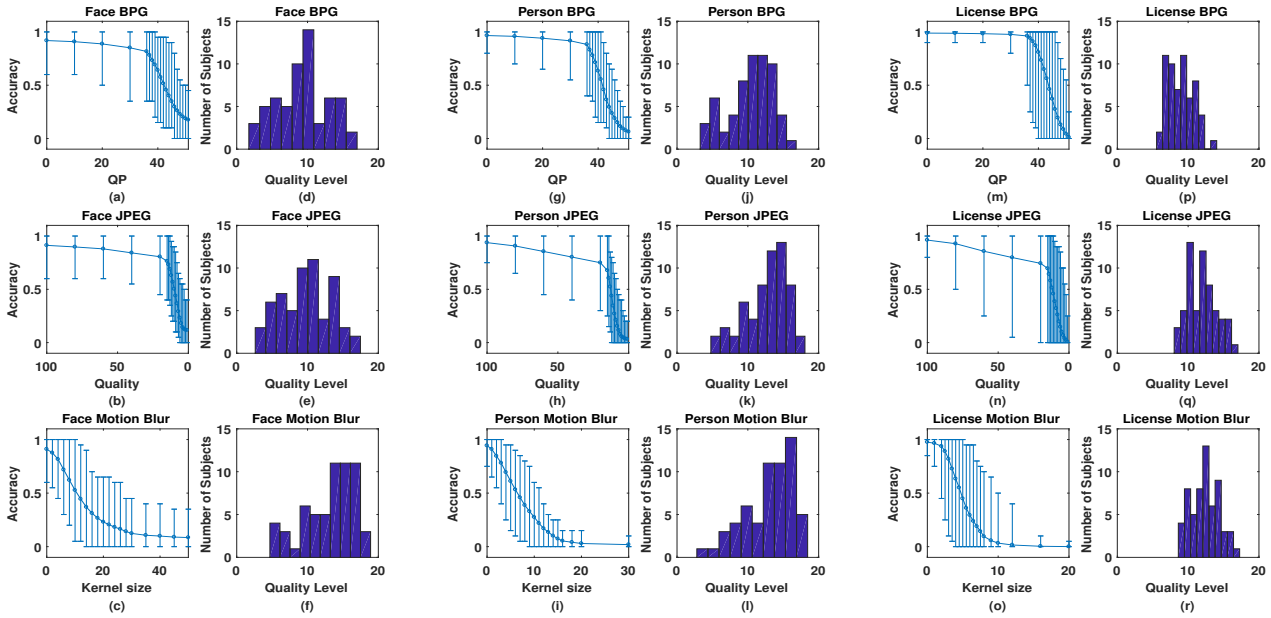

Fig. 3: Subjective recognition accuracy and threshold histograms for different distortions and subsets. (a-c) (g-i) (m-o): Average, maximum and minimum subjective recognition accuracy. (d-f) (j-l) (p-r): Histograms of *subjective recognition threshold*

subjective recognition threshold. Then, average thresholds of each person for each subset and distortion type are calculated. Finally, histograms for average recognition thresholds of valid subjects are obtained. Some conclusions can be drawn from the histograms. Human can recognize clear images correctly, however, tolerance for distortion varies from subject to subject when distortion is introduced. The recognition threshold distribution in the histograms approximate to Gaussian distribution, with data mainly centralizing on moderate distortion levels. A few subjects can recognize correctly even if the distortion is severe, or cannot recognize until the images are clear.

C. Differences between Human and Deep Learning Models

In recent years, artificial intelligence shows extraordinary ability in image recognition, generation, etc. But can artificial intelligence really surpass human or what are the differences between humans and models? We investigate the question based on the proposed database.

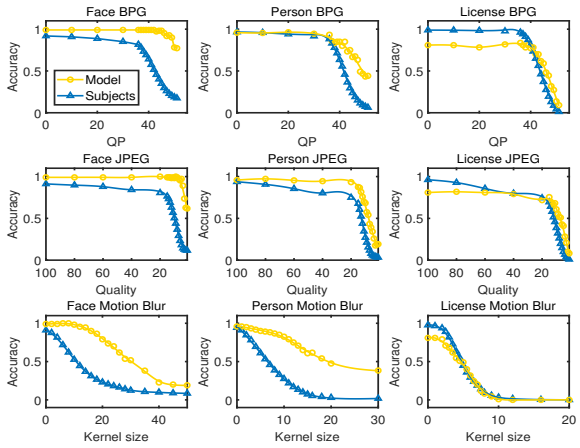

Fig. 4: Average recognition accuracy of humans and models

For each recognition task, we choose a state-of-art model as the representative of AI, namely FaceNet [8], HyperLPR [23] and Deep Person Re-Id [30]. The average recognition accuracies for subjects and models are presented in Fig.4. We can see that for face recognition and person re-identification, deep neural networks surpass human by a large margin when the distortion is severe and for license plate recognition, the superiority is smaller. It means neural networks can better deal with images with bad quality and give the right answer. Furthermore, the drop points of models usually lag behind those of subjects, which indicates neural networks have more tolerance to image distortion. Also, as learned from subplots for motion blur, neural networks are better at deconvolution than human beings.

In Fig.4, the average accuracy of model for reference license plates is lower than human. It can be explained that face and pedestrian images are chosen from large public datasets serving identification tasks so the state-of-art models have already achieved high accuracies on these images. However, the license plate image database is self-built so the CNN model may have difficulty with generalization problems. Another potential weakness of these models is instability. As the distortion increases, the accuracy of models fluctuates occasionally, whereas the accuracy of subjects decreases steadily.

D. Performance Evaluation of IQA Metrics

In a sense, recognition accuracy is similar to image quality as they are both high for reference images and low for distorted images. As a result, we evaluate some well-known objective image quality assessment metrics with Linear Correlation Coefficient (LCC) and Spearman Rank Correlation Coefficient (SROCC) on our database. Higher coefficient means a better correlation with subjective quality judgment.

The performance evaluation shown in Table II demonstrates that existing metrics achieve better results on the pedestrian subset, while seemingly not very effective for license plates. A metric suitable for evaluation should at least yield correlation coefficients higher than 0.9 [31]. However, we can learn from the table that current IQA metrics are not appropriate to predict the recognition accuracy of distorted images and it is possible to propose a specific metric for measuring semantic distortion.

TABLE II: LCC and SROCC performance of objective image quality metrics on our database

	Face		Pedestrian		License plate		All	
	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC
PSNR	0.7878	0.8061	0.8152	0.8505	0.7662	0.7660	0.7429	0.7622
SSIM [32]	0.8476	0.8586	0.8862	0.9060	0.7655	0.7570	0.7975	0.8188
VIF [33]	0.8720	0.8792	0.9112	0.9289	0.7168	0.7119	0.7946	0.8180
IFC [34]	0.8442	0.8515	0.9118	0.9248	0.6384	0.6303	0.7355	0.7514
FSIM [35]	0.8417	0.8549	0.8791	0.9041	0.8043	0.7979	0.8167	0.8031
BRISQUE [36]	0.7405	0.7511	0.6891	0.7081	0.4572	0.4681	0.5631	0.5849

IV. Conclusion

At an attempt to study how human and AI perceive semantic distortion differently, we picked 3 specific targets (face, pedestrian, license plate) and 3 distortion types (JPEG compression, BPG compression, motion blur) based on the surveillance scenario, and measure the semantic distortion with the accuracy of corresponding recognition tasks. We then propose the Semantic Database by collecting the subjective recognition accuracy data. Results show that AIs have a remarkable advantage over human against high distortion, while for low distortion the situations differ from task to task. It is also indicated that current IQA methods are incapable of evaluating quality from the semantic perspective. Variations among human participants and weaknesses of AI such as instability and generalization ability are also discussed.

In the future, the effect of different distortion including compression artifacts, additive noise like rain, haze and adversarial examples would be a key point. For compression artifacts, artificial intelligence is a notch above human. However, it may be a totally different situation for adversarial examples. Since these images are generated to fool machine learning systems, human can easily recognize them but AI fails. Meanwhile, it is urgent to put forward a new metric to measure the semantic distortion for intelligent equipment.

REFERENCES

[1] Y. Pei, Y. Huang, Q. Zou, Y. Lu, and S. Wang, "Does haze removal help cnn-based image classification?" in *European Conference on Computer Vision*. Springer, 2018, pp. 697–712.

[2] H. Sheikh, "Live image quality assessment database release 2," *http://live. ece. utexas. edu/research/quality*, 2005.

[3] N. Ponomarenko *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[4] E. C. Larson and D. Chandler, "Categorical image quality (csiq) database," 2010.

[5] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," *arXiv:1502.00873 [cs]*, Feb. 2015, arXiv: 1502.00873. [Online]. Available: http://arxiv.org/abs/1502.00873

[7] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L2 Hypersphere Embedding for Face Verification," in *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*. Mountain View, California, USA: ACM Press, 2017, pp. 1041–1049. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3123266.3123359

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, Jun. 2015, arXiv: 1503.03832. [Online]. Available: http://arxiv.org/abs/1503.03832

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Procedings of the British Machine Vision Conference 2015*. Swansea: British Machine Vision Association, 2015, pp. 41.1–41.12. [Online]. Available: http://www.bmva.org/bmvc/2015/papers/paper041/index.html

[10] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *LNCS*, vol. 9911, Oct. 2016, pp. 499–515.

[11] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," *arXiv:1704.08063 [cs]*, Apr. 2017, arXiv: 1704.08063. [Online]. Available: http://arxiv.org/abs/1704.08063

[12] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," *arXiv preprint arXiv:1801.09414*, 2018.

[13] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *arXiv:1801.07698 [cs]*, Jan. 2018, arXiv: 1801.07698. [Online]. Available: http://arxiv.org/abs/1801.07698

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, http://vis-www.cs.umass.edu/lfw/results.html#Human, Tech. Rep. 07-49, October 2007.

[15] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, and N. Komodakis, "A Robust and Efficient Approach to License Plate Detection," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1102–1114, Mar. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7752971/

[16] S. Yu, B. Li, Q. Zhang, C. Liu, and M. Q.-H. Meng, "A novel license plate location method based on wavelet transform and emd analysis," *Pattern Recognition*, vol. 48, no. 1, pp. 114–125, 2015.

[17] A. H. Ashtari, M. J. Nordin, and M. Fathy, "An iranian license plate recognition system based on color features," *IEEE transactions on intelligent transportation systems*, vol. 15, no. 4, pp. 1690–1705, 2014.

[18] D. F. Llorca *et al.*, "Two-camera based accurate vehicle speed measurement using average speed at a fixed point," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 2533–2538.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[20] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," *arXiv preprint arXiv:1312.6082*, 2013.

[21] S. Zherzdev and A. Gruzdev, "LPRNet: License Plate Recognition via Deep Neural Networks," *arXiv:1806.10447 [cs]*, Jun. 2018, arXiv: 1806.10447. [Online]. Available: http://arxiv.org/abs/1806.10447

[22] "Openalpr," http://www.openalpr.com.

[23] "Hyperlpr," https://github.com/zeusees/HyperLPR.

[24] "Toward End-to-End Car License Plate Detection and Recognition With Deep Neural Networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8424450/

[25] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 791–808.

[26] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1335–1344.

[27] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1239–1248.

[28] Y. Chen, X. Zhu, W. Zheng, and J. Lai, "Person Re-Identification by Camera Correlation Aware Feature Augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Computer Vision, IEEE International Conference on*, 2015.

[30] "Deep person re-id project," https://github.com/KaiyangZhou/deep-person-reid.

[31] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[33] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–709.

[34] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[35] Z. Lin, Z. Lei, M. Xuanqin, and Z. David, "Fsim: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 20, no. 8, p. 2378, 2011.

[36] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 21, no. 12, p. 4695, 2012.