On the Impact of Lossy Image and Video Compression on the Performance of Deep Convolutional Neural Network Architectures

Matt Poyser, Amir Atapour-Abarghouei, Toby P. Breckon
Department of Computer Science
Durham University, UK.

***Abstract*—Recent advances in generalized image understanding have seen a surge in the use of deep convolutional neural networks (CNN) across a broad range of image-based detection, classification and prediction tasks. Whilst the reported performance of these approaches is impressive, this study investigates the hitherto unapproached question of the impact of commonplace image and video compression techniques on the performance of such deep learning architectures. Focusing on the JPEG and H.264 (MPEG-4 AVC) as a representative proxy for contemporary lossy image/video compression techniques that are in common use within network-connected image/video devices and infrastructure, we examine the impact on performance across five discrete tasks: human pose estimation, semantic segmentation, object detection, action recognition, and monocular depth estimation. As such, within this study we include a variety of network architectures and domains spanning end-to-end convolution, encoder-decoder, region-based CNN (R-CNN), dual-stream, and generative adversarial networks (GAN). Our results show a non-linear and non-uniform relationship between network performance and the level of lossy compression applied. Notably, performance decreases significantly below a JPEG quality (quantization) level of 15% and a H.264 Constant Rate Factor (CRF) of 40. However, retraining said architectures on pre-compressed imagery conversely recovers network performance by up to 78.4% in some cases. Furthermore, there is a correlation between architectures employing an encoder-decoder pipeline and those that demonstrate resilience to lossy image compression. The characteristics of the relationship between input compression to output task performance can be used to inform design decisions within future image/video devices and infrastructure.**

I. Introduction

Image compression is in *de facto* use within environments relying upon efficient image and video transmission and storage such as security surveillance systems within our transportation infrastructure and our daily use of mobile devices. However, the use of the commonplace lossy compression techniques, such as JPEG [1] and MPEG [2] to lower the storage/transmission overheads for such smart cameras leads to reduced image quality that is either noticeable or commonly undetectable to the human observer. With the recent rise of deep convolutional neural networks (CNN [3], [4]) for video analytics across a broad range of image-based detection applications, a primary consideration for classification and prediction tasks is the empirical trade-off between the performance of these approaches and the level of lossy compression that can be afforded within such practical system deployments (for storage/transmission).

This is of particular interest as CNN are themselves known to contain lossy compression architectures - removing redundant image information to facilitate both effective feature extraction and retaining an ability for full or partial image reconstruction from their internals [3], [4].

Prior work on this topic [5]–[8] largely focuses on the use of compressed imagery within the train and test cycle of deep neural network development for specific tasks. However, relatively few studies investigate the impact upon CNN task performance with respect to differing levels of compression applied to the input imagery at inference (deployment) time.

In this paper we investigate whether (a) existing pre-trained CNN models exhibit linear degradation in performance as image quality is impacted by the use of lossy compression and (b) whether training CNN models on such compressed imagery thus improves performance under such conditions. In contrast to prior work topic [5]–[8], we investigate these aspects across multiple CNN architectures and domains spanning segmentation (SegNet, [9]), human pose estimation (OpenPose, [10]), object recognition (R-CNN, [11]), human action recognition (dual-stream, [12]), and depth estimation (GAN, [13]). Furthermore, we determine within which domains compression is most impactful to performance and thus where image quality is most pertinent to deployable CNN model performance.

II. Prior Work

Overall, prior work in this area is limited in scope and diversity [5]–[8]. Dodge et al. [5] analyze the performance of now seminal CNN image classification architectures (AlexNet [14], VGG [15] and InceptionV1 [16]) performance under JPEG [1] compression and other distortion methods. They find that these architectures are resilient to compression artifacts (performance drops only for JPEG quality < 10) and contrast changes, but under-perform when noise and blur are introduced.

Similarly, Zanjani et al. [17] consider the impact of JPEG 2000 compression [18] on CNN, and whether retraining the network on lossy compressed imagery would afford better resultant model performance. They identify similar performance from the retrained model on higher quality images but are able to achieve up to as much as 59% performance increase on low quality images.

Rather than image compression, Yeo et al. [6] compare different block sizes and group-of-pictures (GOP) sizes within MPEG [2] compression against Human Action Recognition (HAR). They determine that both smaller blocks and smaller groups increase performance. Furthermore, B frames introduce propagation errors in computing block texture, and should be avoided within the compression process. Tom et al. [19] add that there is a near-linear relationship between HAR performance and the number of motion vectors (MV) corrupted within H.264 [20] video data, with performance levelling off when 75% of MV are corrupted. Klare and Burge [7], however, demonstrate that there is a non-linear relationship between face recognition performance and bit rate within H.264 video data, with sudden performance degradation around 128kbps (CRF). These contrasting results therefore demonstrate the need to investigate compression quality across multiple challenge domains, whose respective model architectures might have different resilience to lossy compression artifacts.

Multiple authors have developed impressive architectures trained on compressed data, indicating both the potential and need for in-depth investigation within the compressed domain. Zhuang and Lai [8] demonstrate that acceptable face detection performance can be obtained from H.264 video data, while Wang and Chang [21] use the DCT coefficients from MPEG compression [2] to directly locate face regions. The same authors even achieve accurate face tracking results in [22], still within the compressed video domain. The question is evidently:- by *how much* can data be compressed?

These limited studies open the door only slightly on this very question - *what is generalized impact of compression on varying deep neural network architectures?* Here we consider multiple CNN variants spanning region-based, encoder-decoder and GAN architectures in addition to a wide range of target tasks spanning both discrete and regressive outputs. From our observations, we aim to form generalized conclusions on the hitherto unknown relationship between (lossy) image input to target function outputs within the domain of contemporary CNN approaches.

III. Methodology

To determine how much lossy image compression is viable within CNN architectures before performance is significantly impacted we must study a range of second generation tasks, beyond simple and holistic image classification, requiring more complex CNN output granularity. We examine five CNN architectural variants across five different challenge domains, emulating the dataset and evaluation metrics characterized in their respective originating study in each case as closely as possible. Inference models processing images were tested six times, with a JPEG quality parameter in the set $\{5, 10, 15, 50, 75, 95\}$, while video-based models were tested with H.264 CRF compression parameters in the set $\{23, 25, 30, 40, 50\}$. Each model is then retrained with imagery compressed at each of the five higher levels of lossy compression to determine whether resilience to compression could be improved, and how much compression we can afford before a significant impact on performance is observed. Our methodology for each of our representative challenge domains is outlined in the following sections:- semantic segmentation (Section: III-A), depth estimation (Section: III-B), object detection (Section: III-C), human pose estimation (Section: III-D), and human action recognition (Section: III-E).

A. Semantic Segmentation

Pixel-wise Segmantic segmentation involves assigning each pixel in an image (Fig. 1A, above) its respective class label (Fig. 1A, below). SegNet [9] uses an encoder-decoder neural network architecture followed by a pixel-wise classification layer to approach this challenge.

Implementing SegNet from [23], we evaluate global accuracy (percentage of pixels correctly classified), mean class accuracy (mean prediction accuracy over each class), and mean intersection over union (mIoU) against compressed imagery from the Cityscapes dataset [24]. When retraining the network, we use 33000 epochs, with a batch size of 12, fixed learning rate (η) of 0.1, and momentum (β) of 0.9.

B. Depth Estimation

In order to evaluate GAN architecture performance under compression, we need a task decoupled from reconstructing high quality output, to which compression would be clearly detrimental. One such example is computing the depth map of a scene (Fig. 2A, below) from monocular image sequences (Fig. 2A, above).

Using a simplified network from [13], we evaluate RMSE performance of the GAN against the Synthia dataset presented in [25]. We employ $\eta = 0.0001$ and batch size 10 over 10 epochs.

C. Object Detection

In object detection, we must locate and classify foreground objects within a scene (as opposed to semantic segmentation, which classifies each pixel), and compute the confidence of each classification (Fig. 3A). We evaluate mAP of the Detectron FasterRCNN [11] implementation [26] against the Pascal VOC 2007 dataset [27], over mIoU with threshold 0.5:0.95. When training the network, we use $\eta = 0.001$ and weight decay of 0.0005 over 60000 epochs.

D. Human Pose Estimation

Human Pose Estimation involves computing (and overlaying) the skeletal position of people detected within a scene (Fig. 4A). Recent work uses part affinity fields to map body parts to individuals, thus distinguishing between visually similar features.

Using OpenPose [10] we compute the skeletal overlay of detected people in images from the COCO dataset [28]. We evaluate with mean average precision (mAP), over 10 object key-point similarity (OKS) thresholds, where OKS represents IoU scaled over person size. When retraining the network, we use $\eta = 0.001$, and a batch size of 8 over 40 epochs.

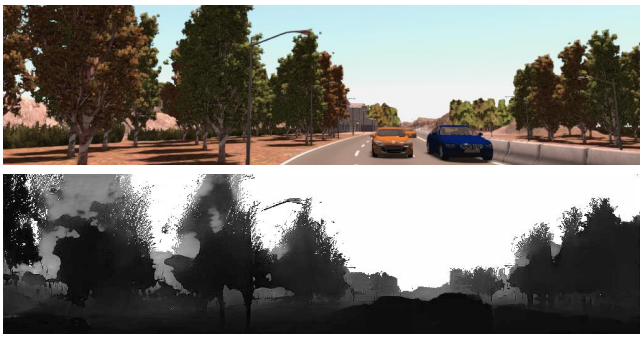
Fig. 1. Results of pre-trained SegNet model [9] on a JPEG image under different compression levels (original RGB image above, computed segmentation map below)

(A) JPEG compression level: 95

(B) JPEG compression level: 15

(C) JPEG compression level: 10

Fig. 2. Results of pre-trained GAN model on a JPEG image under different compression levels (RGB image above, computed depth map below)



(A) JPEG compression level: 95

(B) JPEG compression level: 15

(C) JPEG compression level: 10

E. Human Action Recognition

To classify a single human action - from a handstand to knitting - with a reasonable level of accuracy, we must inspect spatial information from each frame, and temporal information across the entire video sequence.

We implement the dual-stream model from [12]; recognising human activity by fusing spatial and temporal predictions from the UCF101 video dataset presented in [29] (see Fig. 5 for example frames, dramatically deteriorating in quality as H.264 CRF value is increased). To train the temporal stream, we pass 20 frames randomly sampled from the pre-computed stack of optical flow images. Across both streams, we use a batch size of 12, $\beta = 0.9$, and $\eta = 0.001$ for 500 epochs.

Fig. 3. Results of pre-trained FasterRCNN model [11] on a JPEG image under different compression levels

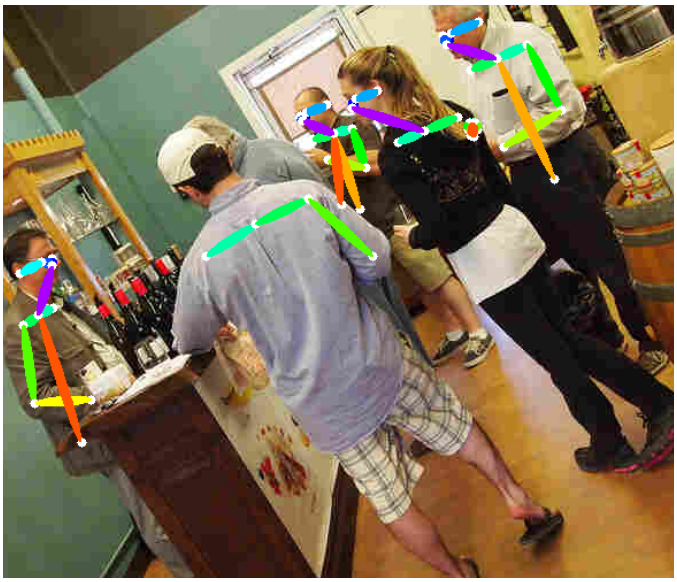
(A) JPEG compression level: 95

(B) JPEG compression level: 15

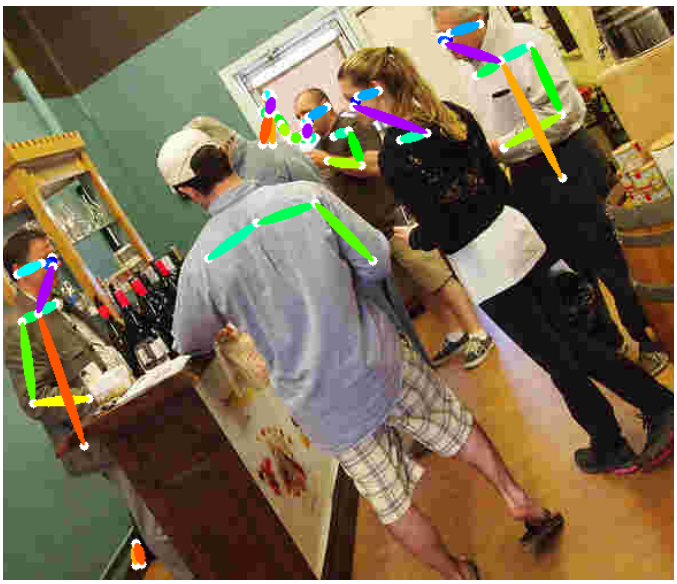
(C) JPEG compression level: 10

Fig. 4. Results of pre-trained OpenPose model [10] on a JPEG image under different compression levels

(A) JPEG compression level: 95



(B) JPEG compression level: 15



(C) JPEG compression level: 10

Fig. 5. One frame taken from a video input to the Two-Stream CNN model [12] under different H.264 compression rates

(A) H.264 CRF value 23

(B) H.264 CRF value 30

(C) H.264 CRF value 40

TABLE I
SEGMENTATION: GLOBAL ACCURACY, MEAN CLASS ACCURACY AND MIOU AT VARYING COMPRESSION RATES

Compression Rate	global ACC	mean ACC	mIoU
95	0.911	0.536	0.454
75	0.909	0.530	0.448
50	0.904	0.523	0.438
15	0.814	0.459	0.338
10	0.794	0.421	0.304
5	0.782	0.364	0.265

(A) after testing a pre-trained SegNet model [9] on compressed imagery

Compression Rate	global ACC	mean ACC	mIoU
95	0.911	0.536	0.454
75	0.910	0.522	0.446
50	0.908	0.503	0.431
15	0.902	0.494	0.420
10	0.895	0.477	0.405
5	0.879	0.445	0.374

(B) after retraining a SegNet model [9] with compressed imagery

TABLE II
DEPTH ESTIMATION: ABSOLUTE RELATIVE, SQUARED RELATIVE, AND ROOT MEAN SQUARED ERROR AT VARYING COMPRESSION RATES (LOWER, BETTER)

Compression Rate	Abs. Rel.	Sq. Rel.	RMSE
95	0.0112	0.0039	0.0588
75	0.0116	0.0039	0.0589
50	0.0123	0.0038	0.0587
15	0.0146	0.0040	0.0599
10	0.0192	0.0042	0.0617
5	0.0283	0.0060	0.0749

(A) after testing a pre-trained GAN model for monocular depth estimation [13] on compressed imagery

Compression Rate	Abs. Rel.	Sq. Rel.	RMSE
95	0.0112	0.0039	0.0588
75	0.0113	0.0035	0.0560
50	0.0103	0.0029	0.0502
15	0.0121	0.0034	0.0556
10	0.0152	0.0031	0.0528
5	0.0159	0.0040	0.0599

(B) retraining a GAN model for monocular depth estimation [13] with compressed imagery

TABLE III
OBJECT DETECTION: MEAN AVERAGE PRECISION AT VARYING COMPRESSION RATES

Compression Rate	mAP
95	0.703
75	0.686
50	0.666
15	0.545
10	0.442
5	0.187

(A) after testing a pre-trained FasterRCNN model [11] on compressed imagery

Compression Rate	mAP
95	0.703
75	0.694
50	0.692
15	0.647
10	0.627
5	0.559

(B) retraining a FasterRCNN model [11] with compressed imagery

IV. Evaluation

In this section, we contrast the performance of the considered CNN architectures under their respective evaluation metrics before and after retraining. From this, we can determine how much we can safely compress the imagery while maintaining acceptable performance. We then propose possible explanations for the variations in resilience of the network architectures to image compression.

A. Semantic Segmentation

From results presented in Table I we can observe that the impact of lossy compression (Table IA) is minimal, indicating high resilience to compression within the network. At the highest (most compressed) compression level, we see global accuracy reduce by 14%, down to 78.2%, while affording 95% less storage cost on average per input image. However, at these heaviest compression rates, the compression artifacts introduced can lead to false labelling. This is particularly prominent where there are varying levels of lighting, affecting even plain roads (Fig. 1C). Subsequently, from Table IB we can see that retraining the network further minimizes performance loss, especially minimizing false labelling of regions. At a JPEG compression level of 5, performance loss is reduced to 3.5%, resulting in global accuracy narrowly dropping below 0.9. Such resilience may stem from the up-sampling by the pooling layers within the decoder pipeline, which are innately capable of recovering information that has been lost during compression, but further investigation is left to future work.

B. Depth Estimation

Analyzing the results in Table II, it is evident that lossy compression markedly diminishes RMSE performance of depth estimation when heavy compression rates are employed (Table IIA). At a JPEG compression level of 15, RMSE has not increased by more than 1.9%, but at a JPEG compression level of 10 and lower, performance begins to dramatically decline (in keeping with that of [5]). However, by retraining the network at the same compression level that is employed during testing (Table IIB), performance loss can be thoroughly constrained. Even at a JPEG compression level of 5, RMSE can be constrained to under 0.0600, improving performance by as much as 20% over the pre-trained network. Other performance measures demonstrate the same trend.

This performance is surprising: we might expect that RMSE would increase (thus lowering performance) after training on compressed imagery, since the GAN generates low quality imagery as the textures and features used to calculate depth estimation are lost, and is therefore unable to improve depth estimation performance. It is possible that it exceeds our expectation due to the encoder-decoder pipeline within the estimation process, which is also employed in the SegNet architecture, and thereby shares its compression resilience.

C. Object Detection

From Table III, we can again discern that performance degrades rapidly at high lossy compression levels (JPEG compression level of 15 or less, see Table IIIA). Applying a JPEG compression level of 15 leads to a 22.5% drop, down to mAP of 0.545, while a JPEG compression level of 5 causes mAP to drop by as much as 73.4%. Furthermore, with higher compression rates, fewer objects are detected, and their classification confidence also falls (Fig. 3C). Their classification accuracy remains unhindered, however. When the network is retrained on imagery lossily compressed at the same level, performance is noticeably improved (Table IIIB). The performance drop as compression rate is increased is delayed from a JPEG compression level of 15 to a JPEG compression level of 5. In fact, the retrained network is able to maintain an mAP above 0.6 even at a JPEG compression level of 10; reducing performance degradation to only 10.8%, while affording a lossy compression rate almost 10-fold higher in terms of reduced image storage requirements.

TABLE IV
HUMAN POSE ESTIMATION: Mean average precision at varying compression rates

Compression Rate	mAP
95	0.711
75	0.689
50	0.655
15	0.413
10	0.323
5	0.098

(A) after testing a pre-trained OpenPose model [10] on compressed imagery

Compression Rate	mAP
95	0.711
75	0.708
50	0.678
15	0.654
10	0.597
5	0.454

(B) after retraining an OpenPose model [10] with compressed imagery

D. Human Pose Estimation

Results in Table IV once again illustrate that lossy image compression (Table IVA) dramatically impacts performance at high rates. Similar to object detection, performance considerably lowers at 15% compression rate, in this case with performance falling by 41.9% to 0.413 mAP. Qualitatively, the network computes precisely located skeletal positions at higher compression rates, but detects and locates fewer joints (Fig. 4B). With high levels of compression (Fig. 4C), the false positive rate increases, and limbs are falsely detected and located. It is likely that optimizing the detection confidence threshold required of joints before computing their location, and thereby maximizing limb detection while minimizing false positives increases performance, especially during high compression. With a retrained network (Table IVB), a compression rate of 15% can be safely achieved before performance degradation exceeds 10%.

While impressive, the results are relatively insubstantial compared to those of other architectures, such as SegNet (Section IV-A, Table I). The difference can perhaps be attributed to the double prediction task within the pose estimation network. Inaccuracies stemming from the lower quality images are not just propagated but multiplied through the network, as the architecture must simultaneously predict both detection confidence maps and the affinity fields for association encodings.

TABLE V
HUMAN ACTION RECOGNITION: TOP-1 ACCURACY FOR EACH STREAM AT VARYING COMPRESSION RATES

Compression Rate	Top-1 Spatial	Top-1 Motion	Top-1 Fusion
23	78.8736	70.1198	83.5485
25	78.7999	44.9225	73.6030
30	78.4563	37.3598	72.2329
40	74.5704	38.9565	70.8803
50	44.1977	15.3267	41.4777

(A) after testing a pre-trained HAR model [12] on video data with varying H.264 CRF encoding values

Compression Rate	Top-1 Spatial	Top-1 Motion	Top-1 Fusion
23	78.8736	70.1198	83.5485
25	78.9056	39.7192	71.7616
30	78.5620	34.3161	70.5765
40	75.9450	9.2550	67.1227
50	62.5165	6.7300	56.2279

(B) after retraining a HAR model [12] with on video data with varying H.264 CRF encoding values

E. Human Action Recognition

From results presented in Table V, it is evident that the impact of lossy compression (Table VA) dramatically increases when we apply CRF factor 50. Conversely to all other examined architectures, we can see from Table VB that retraining the network in fact *decreases* performance.

At first glance, we might expect similar performance to pose detection as with the two stream network for human action recognition, as the errors introduced by compression artifacts propagate through both streams in the network. However, the spatial and motion streams are not trained in tandem. While the spatial stream remains resilient, once again due to the upsampling within the architecture (Section IV-A), the motion stream is almost entirely unable to learn from compressed imagery. As such, retraining the network on compressed imagery in fact reduces overall performance (aside from when using CRF 50, as the spatial stream improvement outweighs the motion stream degradation). Future work may reveal whether better performance might be achieved by retraining just the spatial stream network on compressed imagery, and fusing its predictions with a motion stream trained only on uncompressed imagery.

V. CONCLUSION

This study has investigated the impact of lossy image compression on a multitude of existing deep CNN architectures. We have considered how much compression can be achieved while maintaining acceptable performance, and to what extent performance degradation can be ameliorated by retraining the networks with compressed imagery.

Across all challenges, retraining the network on compressed imagery recovers performance to a certain degree. This study has brought to attention in particular, however, that in very prevalent and so far unexamined network architectures, we can afford to compress imagery at extremely high rates. Segmentation and depth estimation in particular demonstrate resilience against even very significant compression, both by employing an encoder-decoder pipeline. By using retrained models, compression can safely reach as high as 85% across all domains. In doing so, current storage costs can be markedly diminished before performance is noticeably impacted. Hyper parameter optimization of the retrained model can assumedly capitalize on this even further, and in certain domains, such as segmentation, we can already afford to reduce to a twentieth of the original storage cost. It should be noted however, that even a 1 or 2% performance loss may be unacceptable in safety critical operations, such as depth estimation for vehicular visual odometry.

We can further suggest that lossy image compression is potentially viable as a data augmentation technique within RCNN [11] and pose estimation [10] architectures, which receive only mild performance degradation. Networks employing an encoder-decoder architecture (SegNet [9], GAN [13]) would only notably benefit from very significant levels of image compression for data augmentation. However, human action recognition networks, or sub-networks in the case of the two stream approach [12], that consider motion input will not readily benefit from image compression as a data augmentation technique, since they appear unable to learn under such training conditions.

Future work will investigate whether performance is improved by retraining the network with more heavily or lightly compressed imagery than at testing, or even a variety of compression levels. Furthermore, evaluating performance of compressed networks such as MobileNet [30] against compressed imagery would be pertinent, as such light network architectures are prevalent amidst compressed imagery application domains.

VI. ACKNOWLEDGEMENTS

This work was supported by Durham University, the European Regional Development Fund Intensive Industrial Innovation Grant No. 25R17P01847

REFERENCES

[1] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, Apr. 1991. [Online]. Available: http://doi.acm.org/10.1145/103085.103089

[2] D. Le Gall, "MPEG: A video compression standard for multimedia applications," *Commun. ACM*, vol. 34, no. 4, pp. 46–58, Apr. 1991. [Online]. Available: http://doi.acm.org/10.1145/103085.103090

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems - Volume 1*. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–44, 05 2015.

[5] S. F. Dodge and L. J. Karam, "Understanding how image quality affects deep neural networks," *Computing Research Repository*, vol. abs/1604.04004, 2016.

[6] C. Yeo, P. Ahammad, K. Ramchandran, and S. S. Sastry, "High-speed action recognition and localization in compressed domain videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1006–1015, Aug 2008.

[7] B. Klare and M. Burge, "Assessment of H.264 video compression on automated face recognition performance in surveillance and mobile video scenarios," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 7667, Apr 2010.

[8] S.-S. Zhuang and S.-H. Lai, "Face detection directly from H.264 compressed video with convolutional neural network," in *International Conference on Image Processing*, Dec 2009, pp. 2485 – 2488.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *Computing Research Repository*, vol. abs/1511.00561, 2015.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1302–1310.

[11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *Computing Research Repository*, vol. abs/1506.01497, 2015.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 568–576.

[13] A. Atapour-Abarghouei and T. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation," in *Proc. Computer Vision and Pattern Recognition*. IEEE, June 2018, pp. 1–8.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems - Volume 1*. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, p. arXiv:1409.1556, Sep 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Computing Research Repository*, vol. abs/1409.4842, 2014.

[17] F. G. Zanjani, S. Zinger, B. Piepers, S. Mahmoudpour, P. Schelkens, and P. H. N. de With, "Impact of JPEG 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images," *Journal of Medical Imaging*, vol. 6, no. 2, pp. 1 – 9, 2019.

[18] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, Sep. 2001.

[19] M. Tom, R. V. Babu, and R. G. Praveen, "Compressed domain human action recognition in H.264/AVC video streams," *Multimedia Tools Appl.*, vol. 74, no. 21, pp. 9323–9338, Nov. 2015.

[20] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003. [Online]. Available: https://doi.org/10.1109/TCSVT.2003.815165

[21] H. Wang and S.-F. Chang, "A highly efficient system for automatic face region detection in mpeg video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 615–628, Aug 1997.

[22] H. Wang, H. S. Stone, and S.-F. Chang, "FaceTrack: tracking and summarizing faces from compressed video," in *Multimedia Storage and Archiving Systems IV*, S. Panchanathan, S.-F. Chang, and C.-C. J. Kuo, Eds., vol. 3846, International Society for Optics and Photonics. SPIE, 1999, pp. 222 – 234.

[23] M. P. Shah, "Semantic segmentation architectures implemented in pytorch." *https://github.com/meetshah1995/pytorch-semseg*, 2017.

[24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *Computing Research Repository*, vol. abs/1604.01685, 2016.

[25] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Conference on Computer Vision and Pattern Recognition*. IEEE, June 2016, pp. 3234–3243.

[26] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He, "Detectron," https://github.com/facebookresearch/detectron, 2018.

[27] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[28] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *Computing Research Repository*, vol. abs/1405.0312, 2014.

[29] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *Computing Research Repository*, vol. abs/1212.0402, 2012.

[30] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *Computing Research Repository*, vol. abs/1704.04861, 2017.