LEARNT DEEP HYPERPARAMETER SELECTION IN ADVERSARIAL TRAINING FOR COMPRESSED VIDEO ENHANCEMENT WITH PERCEPTUAL CRITIC

Darren Ramsook^{*} Anil Kokaram[†]

SigMedia Group, Trinity College Dublin *ramsookd@tcd.ie, [†]anil.kokaram@tcd.ie

ABSTRACT

Image based Deep Feature Quality Metrics (DFOMs) have been shown to better correlate with subjective perceptual scores over traditional metrics. The fundamental focus of these DFQMs is to exploit internal representations from a large scale classification network as the metric feature space. Previously, no attention has been given to the problem of identifying which layers are most perceptually relevant. In this paper we present a new method for selecting perceptually relevant layers from such a network, based on a neuroscience interpretation of layer behaviour. The selected layers are treated as a hyperparameter to the critic network in a W-GAN. The critic uses the output from these layers in the preliminary stages to extract perceptual information. A video enhancement network is trained adversarially with this critic. Our results show that the introduction of these selected features into the critic yields up to 10% (FID) and 15% (KID) performance increase against other critic networks that do not exploit the idea of optimised feature selection.

Index Terms— Compressed Video Enhancement, Perceptual Optimization, Perceptual Deep Features

1. INTRODUCTION

Lossy compression is necessary to support the current infrastructure of modern media applications. However it introduces a wide range of spatial and temporal artifacts that negatively impact the end user quality of experience [1]. Statistical techniques for artifact removal as a post-processing step has been a long studied research topic [2]. More recently, Deep Neural Networks (DNNs) have been used in different areas of the compression pipeline, from end-to-end learnt compression [3], to purely post-processing [4] and mixed post/preprocessing [5]. The use of DNNs for compressed video enhancement has high potential for its ability to exploit perceptually salient metrics in the training process through the use of generative adversarial networks (GANs), something which is difficult to exploit with traditional statistical methods.

The use of transfer learning for crafting Deep Feature Quality Metrics (DFQMs) has been an emerging area of interest in the Deep Learning Community. Image DFQMs have been shown have a high correlation with human based subjective scores [6]. The underlying features used to generate these scores are the internal representations of high performing pre-trained image classification networks e.g. VGG [7]. LPIPS [8] has been shown to outperform traditional quality metrics such as ℓ_2 , SSIM and FSIM. FloLPIPS [9], an extension of LPIPS that emphasizes areas of motion in a patch, has outperformed other traditional metrics including VMAF and MS-SIM.

While these DFQMs are used in wide range of applications for assessing algorithmic performance and also in the loss-functions for



Fig. 1. Proposed W-GAN setup for compressed video enhancement. The generator take in motion information, the current patch and the patch from the nearest I-frame. The critic network includes selected extracted outputs EfficientNetB3 to assist with restoration.

training networks, the features that are used to generate these new metrics lack explainability. This is a characteristic hallmark of deep neural networks, however there are some approaches that can be used to help segmenting deep features and their importance in the current application.

Contributions of this paper

This paper outlines the use of perceptually relevant deep features in the creation of a neural critic for a Wasserstein-GAN (W-GAN) with gradient penalty [10]. This neural critic is then used for training a compressed video enhancement network using data that is compressed with the H.265 codec. The following contributions are made:

- A method for evaluating deep features of a pre-trained image classification network to isolate feature blocks that have perceptually relevant features.
- The design of a critic network that uses these selected feature blocks as a primary input and an associated compressed video enhancement network that is trained adversarially with the critic.

1.1. Related Work

Perceptually relevant DNN approaches have been recently used within the compression pipeline. Chen et al. [11] investigates the use of loss functions that include perceptual metrics (VMAF, SSIM or MS-SIM) alongside MSE for training an end-to-end DNN image codec. Their results show more than 20% coding gain with VMAF in the loss function when compared to a pure MSE loss. There has been sufficient work that show DFQMs [6, 8, 9] are more closely



Fig. 2. Averaged weights per layer in RDM analysis. Based on this measure of importance, layers 0 and 3 were selected as the top two most significant internal representations. This follows the notion that layers closer to the input of classification networks encode higher perceptual information.

related to human subjective scores. This suggests that using the features that these metrics are dependent on has the potential to yield better results beyond using hand crafted metrics like VMAF. In our proposed model that focuses on post-processing for videos, we utilize extracted prior deep features from EfficientNetB3 directly in the neural critic of our W-GAN for training. We do not optimize for any other traditional metrics or DFQMs directly.

Mohammadi et al. [12] presents an investigation of the perceptual impact of changing the optimization metric of the loss function for a DNN image codec. Results indicate that training with the DFQM DISTS [6] and MS-SSIM yields highest perceptual gain. Images that were compressed with DISTS were selected with the highest winning frequency of 21% when compared to a total of 9 different optimization schemes. In [12], the entire set of deep features is used for the DFQM. In our work we instead identify the subset of features which are most relevant and deploy them in a more efficient adversarial training approach.

Ma et al. [13] presents a post-processing GAN based approach that uses a mixture of multiple objective metrics. Their loss function incorporates the distance between complete deep features of a degraded-reference pair directly. In our approach, we do not use a difference between deep feature pairs. Instead, we improve performance by only using a subset of features that are perceptually relevant. These features are given to our critic network. This allows the critic to learn distinguishing features between deep representations of clean and degraded patches.

2. DEEP FEATURE SELECTION

DFQMs such as LPIPS, FloLPIPS and DISTS are all dependent on the internal representations of pre-trained classification networks. The deep representation of an image is generated by extracting the output of intermediate layers in the network. Features are usually collected at the final output of the convolution filter bank operation of that layer.

These features can be represented by a set of rank 3 arrays whose height and width depends on the size of the input image. The number of channels in the feature array is dependent on the number of convolution filters used at that layer and can be represented by the layer depth. For an arbitrary image input \tilde{I} , the set of generated features F can be represented by $F_z(\tilde{I}) \in F(\tilde{I})$, where z represents the depth at which a feature is accessed.

Historic DFQMs (Q) [6, 8, 9] are based on comparing the internal representations of a reference (x) and degraded (y) patch, Q(F(x), F(y)). A key point is that the F_z are extracted from *frozen* layers of a network trained for classification. Hence not all of the features will have significant relevance for perceptual criteria. Therefore, in our proposed method, our critic utilizes only a subset of perceptually relevant features F^{α} from a single input \hat{x} : $F^{\alpha}(\hat{x}) \subset$ $F(\hat{x})$. Discovering which subset is perceptually relevant is one of our main contributions. The intuition is that there is a set of compression invariant features that can be used by the W-GAN in restoration of a given patch. In addition, by selecting this subset we reduce the dimensionality of Q() and thus improve computational performance of the critic network by allowing to focus on a subset of key features.

A dataset of 99 reference 512×512 patches was used to generate deep feature representations from the classification network EfficientNetB3 [14] pre-trained on ImageNet. This network was chosen for its relatively small number of parameters and increased performance over other classificaton networks. The 99 patches were compressed using the H.265 codecs at 6 linearly spaced constant rate factor (CRF) points in the operating range. This resulted in 99 reference and 594 degraded full feature representations.

Each F for a single patch contains rank-3 arrays that total \approx 3 million elements. Traditional statistical techniques for dimensionality reduction proves intractable for these high element multi-dimensional feature arrays. To overcome this limitation, we take inspiration from neuroscience [15, 16]. In [16], inferior temporal (IT) response to images presented to humans were compared to responses from layers in a pre-trained classification network when presented with those same images. Their goal was to discover which CNN layers were able to best model the human response. To do that, the authors exploited a dissimilarity measure [15] for the CNN layer outputs which was then used to predict the same dissimilarity response from the human stimulus. In our case we are interested in discovering the specific layers in our representation (Efficient-NetB3) which best enable the critic network to discriminate between degraded and clean video. In a sense the stimulus here is the input images and the response is the output of our CNN layers. Hence we follow the same strategy as the neuroscience case.

We first define Representational Dissimilarity Matrices (RDMs) for responses F_z from layer z and F from all layers, as follows.

$$R^{Y}(i,j) = d(\mathbf{F}(y^{i}), \mathbf{F}(y^{j}))$$
(1)

$$R^{cX_z}(i,j) = d\left(F_z(x_c^i), F_z(x_c^j)\right) \tag{2}$$

where $d(;\cdot)$ is the euclidean distance, y_k is the kth patch in our clean dataset and x_c^k is the k th patch in the datset degraded through compression with CRF c. Hence $F_z(x_c^i)$ is the response of the zth layer to the degraded patch i. $R^Y(i, j)$ is therefore a measurement of the auto-dissimilarity between all the clean patches in the dataset, while $R^{cX_z}(i, j)$ is the auto-dissimilarity across the degraded patches.

By building a linear predictor for $R^Y(i, j)$ based on linear combination of elements across layers in $R^{cX_z}(i, j)$, we can select layers based on significance for prediction. The linear predictor is defined through weights $_c\beta_z$ applied to each layer/degradation element in R^{cX_z} as follows.

$$\hat{R}^{Y}(_{c}\beta_{z}) = \sum_{\forall z} \sum_{\forall c} {}_{c}\beta_{z} R^{cX_{z}}, \qquad (3)$$

We estimate ${}_c\beta_z$ by minimising the cosine distance between $\hat{R}^Y(c\beta_z)$ and R^Y . We apply the additional constraint of non-negative weights with the BFGS algorithm[17].



(a) Critic. An input patch is processed by EfficientNetB3 and only layers 0 and 3 (purple) are used in producing output scores. Computation stops at layer 3, other layers are included for clarity.



(b) Generator. Motion magnitude and direction m_t of the current frame is concatenated with the current RGB patch f_t . The corresponding patch from the nearest intra-coded frame is fed into a CNN block (blue). The resulting features are concatenated with the center of the UNet (red). These features are then fed into an attention module (CBAM [18]). The restoration parameter d_t guides the enhancement network and Δ_t is a weighting factor that scales the features extracted from the I-frame.

Fig. 3. Proposed Critic (a) and Generator (b) networks.

By averaging the estimated weights across degradation level c (6 in our case) we estimate the relative importance of each layer. This is shown in Figure 2. Layers 0 and 3 of EfficientNetB3 are therefore chosen for our critic network as they clearly encode most perceptual information.

3. NETWORK ARCHITECTURE

Critic: Our critic network takes a patch as input and produces a feature block that represents the patch degradation level. This is adapted from the PatchGAN architecture [19]. The main contribution of our critic network is the introduction of the selected EfficientNetB3 layers as a primary input. This is shown in Figure 3(a).

The extracted features are processed in parallel by CNN Blocks. The input frame is also processed directly to capture information that EfficientNetB3 may have missed. These are then concatenated and processed further to produce the feature block. There is no activation function in the final layer. This follows the requirement of a W-GAN for the critic output to be unbounded.

To compare our proposed critic with optimal layers (C_{opt}) , we also train with critic networks that:

- 1. Uses the output from layers 5 and 6 of EfficientNetB3 ($C_{5,6}$).
- 2. Does not use the EfficientNetB3 module, and hence the entire critic is fully trainable (C_t) .

Generator: Our generator (Figure 3(b)) uses a UNet structure as the main network. Motion vector information (magnitude and direction) are concatenated with the input RGB patch. This is then fed into a UNet with skip connections. Features from the nearest Intra-coded frame (I-frame) are extracted by a CNN block and is concatenated with the center of the UNet. It is assumed that the nearest I-frame has

Table 1. The impact of varying Λ for C_{opt} compared against other critics and VBM4D.

	Lambda	\mathbf{PSNR}_Y	PSNR _{CBCR}	SSIM	VMAF	LPIPS	FID	KID
	0.0	26.92	33.24	0.80	46.82	0.45	0.44	3.3e-4
\mathbf{C}_{opt}	1.0	26.07	34.82	0.79	43.67	0.45	0.44	3.5e-04
	10.0	27.98	35.06	0.79	50.43	0.44	0.45	3.6e-04
	100.0	28.25	36.87	0.81	48.72	0.43	0.40	3.8e-04
	1000.0	29.11	37.75	0.83	47.28	0.41	0.36	<u>3.2e-04</u>
	10000.0	29.43	38.12	0.83	47.28	0.42	0.37	3.7e-04
$C_{5,6}$	1000.0	29.64	38.41	0.83	45.91	0.43	0.39	3.8e-04
\mathbf{C}_t	1000.0	29.06	37.65	0.82	49.50	0.44	0.40	3.5e-04
VBM4D	-	30.86	<u>39.21</u>	0.82	41.09	0.43	1.2	1.2e-02
Degraded	-							

a high amount of texture related to the current scene. The features extracted from the I-frame are scaled by the distance parameter Δ_t , where $0 \leq \Delta_t \leq 1$. This allows the I-frame features to have a greater weight if they are closer to the current patch. Δ_t is calculated as : $\Delta_t = e^{-0.02(\gamma)}$, where γ is the absolute distance between the frame index of input patch and the nearest I-frame.

The user-defined restoration parameter d_t specifies the strength of the restoration and is proportional to the amount of degradation present in the input patch. This value is fed into all stages of the UNet to guide the restoration.

4. TRAINING

A dataset with 2649 degraded/reference pairs was created by degrading clips using the CRF parameter on an H.265 codec such that the resulting PSNR_Y of the degraded video was \leq 38dB. More details about the dataset can be found in the supplementary material [20]. All training patches were 512x512.

The degradation parameter d_t , which controls the strength of the restoration network, was set to be the LPIPS score between the degraded input patch and the reference patch. Our motion magnitude and direction (m_t) was calculated using the DeepFlow algorithm [21] and the Luma channel of the previous patch, current patch and next patch.

Our generator was initially trained using the squared l_2 norm loss for 25 epochs. The outputs from this initially trained generator was then used to train the critic network for 10 epochs. This type of training is adapted from the NoGAN approach [22]. The Adam optimizier with learning rate 1e-04 was used in both cases. The generator and critic network were then trained adversarially for 50 epochs using Adam optimizers with an initial learning rate of 1e-04 that decayed to 1e-06. Our adversarial training routine follows [10], with the loss function for both the critic and generator as follows.

$$\mathcal{L}_{critic} = -\left[C(y_t) - C(G(f_t))\right] + \lambda \Gamma \tag{4}$$

$$\mathcal{L}_{gen} = -C(G(f_t)) + \Lambda \left(l_2(\hat{f}_t, y_t) \right)$$
(5)

Where C, G are the critic and generator network respectively. λ , Γ are the gradient penalty weight and gradient penalty function (as in [10]). Λ is a weight that allows the generator to strike a compromise between MSE $(l_2(\cdot, \cdot))$, and perceptual criteria. This is important because generative networks like these do have the tendency to synthesise or hallucinate content which *looks like* picture content but is not constrained by the input pixel texture. The MSE weight Λ therefore allows the output to be constrained to be related more closely to the input content.



Fig. 4. Generated patches. VBM4D (b) has the highest PSNR but suffers from oversmoothening. This is a characteristic of optimzing with respect to MSE alone. Our proposed method C_{opt} with $\Lambda = 10$ yields a high detailed image but the color does not match the original material. $C_{5,6}$ has the highest PSNR of all critic approaches, but still has a loss of detail. C_t has higher DFQM scores than $C_{5,6}$, however there is a consistent streaking effect overlaid on degraded parts of the image. C_{opt} with $\Lambda = 1000$ has the best performing DFQM metrics.

5. RESULTS

Table 1 shows results of performance of the various systems tested using a unique test set of 501 reference/degraded clips. In addition to the usual performance metrics of VMAF and PSNR, we report Frechet Inception Distance (FID) and Kernel Inception Distance (KID) [23]. These latter metrics express distance in terms of the output of features from an Inception network. They have recently increased in importance for perceptual evaluation of picture quality as used [24].

In order to select the optimal value of Λ , the table reports performance for C_{opt} under the effect of varying Λ on a log scale. $\Lambda = 1000$ is chosen as optimal given that it minimises KID and FID. Our proposed critic C_{opt} (using perceptual layers) is the best performer with up to 10% (FID) and 15% (KID) average improvement compared to other critics ($C_{5.6}$, C_t).

 Λ is used to weight the effect of the MSE loss function when training the generator. With lower Λ the output content has a high level of detail but the color of the generated patches does not match the source material. This effect can be seen in Figure 4(e) and is corroborated by the low $PSNR_{Y/CBCR}$ scores. As Λ increases the resulting PSNR_{Y/CBCR} also increases as MSE is inversely proportional to PSNR. The other critic configurations $C_{5,6}$ and C_t were also trained with $\Lambda = 1000$. Patches generated from training with $C_{5,6}$ contain less detail than the other critic configurations, however $PSNR_{Y/CBCR}$ was the best performing among critics. This effect can be seen in Figure 4(f). Training with C_t yielded DFQMs that were better performing than $C_{5,6}$, however there are visible streaking effects over areas of high distortion. This can be seen in Figure 4(g). Considering the performance of more traditional statistical approaches, we find our optimal critic to outperform VBM4D in DFOMs. Patches generated from VBM4D suffered from the

"over-smoothing" effect that is characteristic of MSE optimized algorithms. The visual impact of this is shown in Figure 4(b). This validates the need for perceptual criteria.

6. DISCUSSION & CONCLUSION

Drawing upon the principles of DFQMs, such as LPIPS, we developed a critic that employs layers from a pre-trained classification network to enhance critic proficiency. Although our model displays lower performance in conventional metrics such as VMAF, PSNR, and SSIM, it generates output that exhibit a close resemblance to the source content and this is reflected in performance increases in LPIPS, FID and KID (Shown in Table 1) as well as visual evidence.

Our proposed critic C_{opt} , outperforms the other critics which indicates that the use of the selected layers from EfficientNetB3 exploits latent information that is invariant to degradation from compression. The superior perceptual performance achieved by our proposed model, as compared to VBM4D, provides evidence of the capactiy of the critic to focus on highly informative perceptual features beyond the conventional distance metrics.

A drawback of pure adversarial training for compression enhancement purposes is the introduction of hallucination artifacts, which can cause the appearance of plausible image features that are not present at all in the source content. To mitigate this, we use a weighted loss function using Λ to combine both perceptual critic and MSE. Lower values of Λ imply more emphasis on perceptual loss and hence result in pictures with greater detail.

In future, we plan to conduct a subjective study to further validate our approach. We also plan to measure the importance of different layers of the critic and the generator. Data, full network architectures and models are available at [20].

7. REFERENCES

- Liqun Lin, Shiqi Yu, Liping Zhou, Weiling Chen, Tiesong Zhao, and Zhou Wang, "Pea265: Perceptual assessment of video compression artifacts," *IEEE Transactions on Circuits* and Systems for Video Technology, vol. 30, no. 11, pp. 3898– 3910, 2020.
- [2] Mei-Yin Shen and C.-C.Jay Kuo, "Review of postprocessing techniques for compression artifact removal," *Journal of Visual Communication and Image Representation*, vol. 9, no. 1, pp. 2–14, 1998.
- [3] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell, "Full resolution image compression with recurrent neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5306–5314.
- [4] Fan Zhang, Di Ma, Chen Feng, and David R Bull, "Video compression with CNN-based postprocessing," *IEEE MultiMedia*, vol. 28, no. 4, pp. 74–83, 2021.
- [5] Onur G Guleryuz, Philip A Chou, Hugues Hoppe, Danhang Tang, Ruofei Du, Philip Davidson, and Sean Fanello, "Sandwiched image compression: wrapping neural networks around a standard codec," in 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 3757–3761.
- [6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.
- [7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- [8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [9] Duolikun Danier, Fan Zhang, and David Bull, "FloLPIPS: A bespoke video quality metric for frame interpolation," in 2022 Picture Coding Symposium (PCS). IEEE, 2022, pp. 283–287.
- [10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein GANs," *Advances in neural information processing systems*, vol. 30, 2017.
- [11] Li-Heng Chen, Christos G. Bampis, Zhi Li, Andrey Norkin, and Alan C. Bovik, "Perceptually optimizing deep image compression," 2020.
- [12] Shima Mohammadi and João Ascenso, "Perceptual impact of the loss function on deep-learning image coding performance," IEEE, 2022, pp. 37–41.
- [13] Di Ma, Fan Zhang, and David R. Bull, "CVEGAN: A perceptually-inspired GAN for compressed video enhancement," arXiv, 2020.
- [14] Mingxing Tan and Quoc Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [15] Nikolaus Kriegeskorte and Rogier A. Kievit, "Representational geometry: integrating cognition, computation, and the brain," *Trends in Cognitive Sciences*, vol. 17, no. 8, pp. 401– 412, 2013.

- [16] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte, "Deep supervised, but not unsupervised, models may explain it cortical representation," *PLOS Computational Biology*, vol. 10, no. 11, pp. 1–29, 11 2014.
- [17] R. Fletcher, *Newton-Like Methods*, chapter 3, pp. 44–79, John Wiley and Sons, Ltd, 2000.
- [18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision* (ECCV), 2018, pp. 3–19.
- [19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer* vision and pattern recognition, 2017, pp. 1125–1134.
- [20] Darren Ramsook, "Supplementary material," https://github.com/DarrenR96/DeepHyperparameter.git.
- [21] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid, "Deepflow: Large displacement optical flow with deep matching," in 2013 IEEE International Conference on Computer Vision, 2013, pp. 1385–1392.
- [22] Antoine Salmona, Lucía Bouza, and Julie Delon, "Deoldify: A review and implementation of an automatic colorization method," *Image Processing On Line*, vol. 12, pp. 347–368, 2022.
- [23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] Fabian Mentzer, Eirikur Agustsson, Johannes Ballé, David Minnen, Nick Johnston, and George Toderici, "Neural video compression using gans for detail synthesis and propagation," in Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI. Springer, 2022, pp. 562–578.