DV2DM: A Learning-based Visible Difference Predictor for Videos

Anonymous Authors

***Abstract*—Video has emerged as a predominant medium in today's digital era. Efficient storage and transmission of video necessitate encoding and compression, yet excessive compression can significantly degrade quality. Visible Difference Predictor (VDP) metrics are pivotal in predicting and quantifying the visibility of image distortions, relying on models of the human visual system (HVS). While deep learning enables advanced image VDPs, video VDPs encounter greater complexities. In this work, we introduce the Deep Video Visible Difference Metric (DV2DM) and the ViLocVis dataset—the first of their kind for video VDP enhanced by deep learning and inclusive of manually annotated heatmaps. The ViLocVis dataset, a hybrid of manually annotated and synthetic data, accounts for variables including display luminance, frame rate, and viewing distance. Utilizing a calibrated display model coupled with a customized U-Net architecture, our approach captures spatial-temporal features effectively, demonstrating significant performance improvements. We also present a novel pipeline specifically tailored for video VDP challenges, showcasing substantial achievements in this domain.**

***Index Terms*—Visible Difference Predictor, Videos Compression, Human Visual System**

I. Introduction

VISIBLE Difference Predictor (VDP), intending to predict a map of pixel-wise visible difference from image pairs, plays an important role in giving effective suggestions on analyzing difference based on Human Visual System (HVS). With the development of communication technology, especially the rapid popularization of mobile Internet, the application of video in daily life is becoming more and more frequent. The urgent need for video compression means that better metrics need to be found. Such prediction is of vase application space in computer vision areas such as image/video compression, reconstruction, rendering, and Virtual Reality (VR) / Augmented Reality (AR).

Note that VDP is quite different with similar field of research, quality metrics, which aims to evaluate the quality of an image by a single value. On the one hand, VDP considers the human vision system factors hence it is a subjective metric. On the other hand, The VDP algorithm, in contrast, intends to returns a pixel-wise map instead of a single value, which providing with difference details on each location of the image [44], [45].

Most existing VDP work focuses on the image domain. However, with the proliferation of mobile Internet, video has become one of the most popular media today, making the need for video VDP increasingly urgent. For example, the transmission and storage of video almost always involves compression algorithms. In general, compression algorithms strive to reduce the storage space occupied by the video while losing as little picture information as possible. However, classical video compression algorithms often ignore the subjective perception of the viewer, which means over-compression and computational waste on the discernable areas whilst under-compression and space waste on the discernable areas [13].

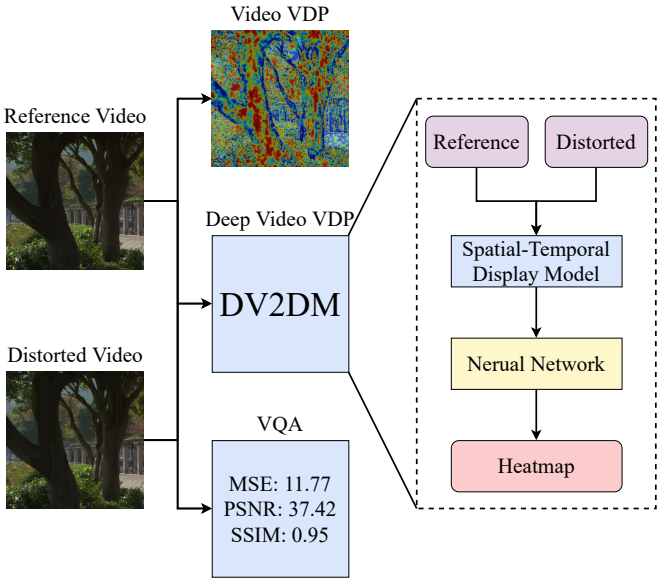

Fig. 1. The framework of our DV2DM, which consists a display model and a nerual network. Different from video quality assessment (VQA), visual difference predictor predicts the visual heatmap instead of a single value.

However, compared to VDP for images, video VDP faces more challenges, such as time dimension and dynamic content. Due to the temporal continuity between frames, the distortion may persist for several frames, affecting the overall perception of the viewer [28]. In addition, it is difficult to cover every frame of the video with annotation information, which makes predicting visible differences in a dynamic scene more complex than in a static image. Different video frame rates can affect the visibility of distortion. For example, frame rates, either too high or too low, might spoil the visibility of differences, for sake that high frame rate may results into less perceptible details and too slow frame rate may shorten the relative window of visibility [40].

In this work, we explore Deep Video Visible Difference Metric (DV2DM), a method and tool for evaluating and predicting visual differences or distortions in video content which can be noticed by an observer, quantifying visual differences and simulating human visual response through algorithms [42].

This work provides the machine learning and computer vision community with a new challenging task, which is significant for the measuring and enhancing of video quality in a common human viewer's sense, especially for this era of video media. The insights for subjective metric on visible difference invokes a wide range comprehensive research interest on human vision system as well as the psychological consideration. Moreover, the publication of our data pave the way for subsequent exploration in this field. Specifically, our main contributions are as follows:

- We propose an effective method to apply deep learning to the Video VDP problem and achieve the state-of-the-art performance. which is based on a U-Net backbone design tailored for VDP. The model architecture utilizes a Siamese U-shaped network architecture with two downsampling branches and only one upsampling branch.
- We consider influences to visible metric from a group of more complicated environment conditions, such as luminance, pixel per degree, and frame rate, which can better handle the real video-watching scenes.
- We provide the subsequent VDP researches with a novel video VDP dataset named ViLocVis, which covers human-annotated data by 10 volunteers on 11 open datasets, and synthetic data derivated using 12 viewing conditions.

II. Related Work

A. Human Visual System and Quality Metrics

1) Human Visual System: The Human Visual System (HVS) refers to the biological system facilitating visual perception, encompassing the eyes, neural pathways, and parts of the brain that process visual data [25]. The study of HVS is important for designing efficient algorithms from the perspective of human eye perception. For example, the HVS is leveraged to design efficient video and image compression algorithms, where information less perceived by the human eye might be deprioritized or omitted during compression [22], [35].

2) Image Quality Metrics: The measurement for quality of an image is quite a mature field of research. Fundamental approaches (e.g. MAE, MSE, PSNR) simply measure the average difference in whole. SSIM [39], MS-SSIM [38], and FSIM [47] further account for the contextual information and structural similarity of image pairs. Machine-learning-based method [5] treats the quality metric problem as a classification process, where the quality is classified into a five-level scale based on the Support Vector Machine, whereas suffering from the subjective definition on quality levels. [48] argues that human perception on similarity depends on high-order structure and is context-dependence, hence introduced a CNN-based method, which is demonstrated to outperform substantially.

Recent advancements in display technologies highlight limitations in existing image quality metrics which assume consistent dynamic ranges between reference and test images. Addressing this, [3] introduced a novel metric that can compare images with varying dynamic ranges, leveraging insights from the human visual system to detect and classify visible structural changes in images. The utility of this metric was further demonstrated in evaluating tone mapping operators and diverse display characteristics.

3) Video Quality Metrics: Video quality is evaluated in different ways from image, due to differences in HVS's handling dynamic and static content. For example, high-frequency noise can be clearly shown in image, whereas less noticeable in high-frame-rate video due to fusion of successive frames [19].

Video Quality Assessment (VQA) techniques are bifurcated into no-reference and full-reference metrics. No-reference metrics evaluate videos relying solely on the test video itself [31] while full-reference metrics compare distortions in the test video against the reference video. Early studies, notably [41] used basic statistical methods to model discrete cosine transform (DCT) quantization noise visibility in digital video quality metrics. [43] discusses the evolution of video quality metrics towards hybrid models blending objective and subjective elements. Machine learning advancements leverage accuracy and consistency [23], [30]. Deep learning further advances with end-to-end neural network [1], [12], [14]. [12] further takes spatio-temporal perception CNN into account. For HDR videos, [24] introduces a calibrated method focusing on noticeable distortions through spatial segmentation and temporal pooling.

B. Visible Difference Predictor

1) VDP: While VQA provides overall video quality assessment through scores, VDP offers finer pixel-level predictions of visible differences in images or videos, which emulates human visual perception to predict the noticeability of visual distortions. It can be traced back to [9] and subsequent research tailors white-box visibility metrics to assess the possibility for human to detect contrasts between image pairs.

The Spatial-CIELAB metric [49] focuses on the color reproduction errors of digital images. It extends the traditional CIELAB [10], which is originally designed for matching large uniform colored areas, to general chromatic images, by tailoring color transformation and spatial filters according to psychophysical experiments on the HVS. Note that, the spatial kernels relies on a Contrast Sensitivity Function (CSF) taking the exponential form, which is over-simplified and cannot generalize well to the more complex cases of images.

Subsequent to early VDP models [9], [18], [27], which incorporated complex low-level vision models but were limited by data scarcity, hindering extensive machine learning application. These models relied on detailed HVS models with minimal trainable parameters, restricting efficacy in complex scenarios. [44] mitigates this by introducing a new dataset for image distortion visibility, enabling machine learning advancements. [45] expands this by incorporating viewing condition factors such as luminance and viewer distance.

With the shift to High Dynamic Range (HDR) imaging, [17] enhances VDP to accurately predict visible differences in HDR scenarios, considering HDR display contrast ratios and real scene adaptation conditions. This approach, calibrated with high-end HDR displays, also accounts for optical light

scattering and local adaptation. Further advancements in HDR-VDP are noted [18], [24].

The trade-off between accuracy and computational efficiency in SDR and HDR metrics remains a challenge. Metrics like HDR-VDP2.2 [24] emulate the HVS well but are computationally demanding. Simpler metrics like PSNR or MSE often miss critical HVS elements. To balance efficiency and accuracy, [4] introduces NoR-VDPNet++, a CNN-based metric optimizing the precision of HDR-VDP2.2.

2) Video VDP: In video contexts, visibility difference prediction necessitates additional considerations like spatio-temporal characteristics and spatial error signals [12]. Despite its significance, research on video VDP remains scarce. [19] pioneers by introducing the first video VDP tailored for foveated displays, such as VR or head-mounted displays, and involves calibration of psychophysical models and the HVS.

III. Data Collection

To collect responses of the human visual system (HVS) to distorted videos under various luminance and viewing distance, we gathered a dataset named ViLocVis, which comprises of several reference-distorted video pairs and corresponding marking videos. The dataset can be divided into two parts: manually annotated data and synthetic data.

TABLE I
Datasets Details

Dataset	Videos	Resolution	FPS
HEVC-B [36]	5	1920x1080	24, 50, 60
HEVC-C [36]	4	832x480	30, 50, 60
HEVC-D [36]	4	416x240	30, 50, 60
HEVC-E [36]	6	1280x720	60
HEVC-F [36]	4	1280x720, 1024x768, 832x480	20, 30, 50
LIVE-APV [32], [33]	45	3840x2160	25, 30
LIVE-SJTU [21]	14	1920x1080	24, 25, 30
MCL-JCV [37]	30	1920x1080	24, 25, 30
MCML [7]	10	3840x2160	30
SJTU4K [34]	15	3840x2160	30
UVG [20]	16	3840x2160	50, 120

The reference videos are from open source datasets, including the HEVC Standard Test Sequences (Class B, C, D, E, and F) [36], LIVE-APV [32], [33], LIVE-SJTU [21], MCL-JCV [37], MCML [7], SJTU4K [34] and the Ultra Video Group (UVG) dataset [20]. Table I provides a summary of the datasets. These videos cover a wide range of resolutions and frame rates in common scenarios.

A. Data Preparation

Before the annotation process, considering that higher video resolutions result in increased storage and computational costs, the videos in the original datasets are downsampled from 3840x2160 to 1920x1080. Additionally, to facilitate the annotation process by enabling annotators to simultaneously view reference and distorted videos, thereby enhancing the ability to discern and annotate distortions in the videos, we cropped the videos into a square format to align with the resolution ratio of the annotation display monitors.

Since the most common distortion in video is caused by compression, we employed compression method to generate distorted videos. considering the popularity of compression

Fig. 2. The user interface of our labelling tool, where the layout is: A) the reference video, B) the distorted video, C) the progress bar, D) Current frame/Total frames, and E) Number of frames have been saved.

Fig. 3. The experiment environment, where the left photo shows the physical environment for subjects to label the videos, and the right photo shows the same environment at the actual ambient lighting conditions.

encoders, we selected the widely used H.264 and H.265. During the compression process, we appropriately adjusted the compression parameters to ensure that most distortions in the compressed videos are perceptible.

It is note-worthy that there are also other common types of distortion, such as noise, blur and watermark. In addition, new distortions introduced by neural networks are also worth attention with respect to the emerging of learning-based compression and reconstruction algorithm [2], [6], [11], [16]. However, as an initial work, we carefully start with the classical compression algorithm and stress out the importance of our first step into locally annotated videos.

B. Manually Annotated Data

In order to understand human's visibility on distortion in videos, we conducted a set of experiments to collect a dataset from participants. Specifically, we recruited 10 subjects to perform video annotation in a unified experimental environment. In Figure 2, our annotation software is shown. The left and right images show the reference video and the distorted video respectively. We asked the participants to identify the difference between the two images. Note that in order to characterize the dynamic nature of the video, we asked subjects to consider not only the differences between static images but also the dynamic differences between consecutive frames in the video in the segment before the annotated frame.

The observers viewed the display at all combinations of three distance levels (40 cm, 65 cm and 80 cm, which correspond to pixel per degree of 30, 50 and 60, respectively) and three screen luminance levels (110 cd/m^2, 200 cd/m^2, and 300 cd/m^2).

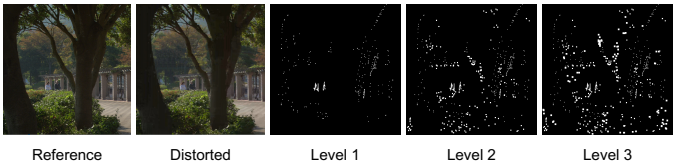

Fig. 4. Different levels of labelling: the "Reference" refers to a reference an original frame from the reference video, the "Distorted" refers to a distorted frame from the compressed video, and {"Level 1", "Level 2", "Level 3"} showcases different levels of labelling from coarse to fine, e.g. luminance from low to high, ceteris paribus.

However, allowing subjects to annotate the distortion frame by frame poses a series of challenges, such as a long period of annotation leading to fatigue and adaption to the distortion level, resulting in the inability to detect certain distortions. To avoid frame-by-frame annotation while obtaining annotated data with high fidelity, we utilized the optical flow method to achieve sparse annotation of videos, where subjects only need to annotate a few frames.

Suppose the reference video is x, the distorted video is $\hat{x}$ and the marking video is y. According to the optical flow method, we use f and g to denote forward and backward optical flow respectively. Due to the temporal continuity of video frames, there should be $x_{i+1} = f(x_i)$ and $x_{i-1} = g(x_i)$, where i is the frame index. Simultaneously, distortions in the video frames are also temporally continuous, thus the probability of subjects annotating the same region is continuous. Therefore, we can obtain $y_{i+1} = f(y_i)$ and $y_{i-1} = g(y_i)$, where f and g can be obtained given the relevant video frames. Furthermore, we can derive $y_{i+1} = f(f(y_{i-1})) = f \circ f(y_{i-1}) = f^2(y_{i-1})$ and $y_{i-1} = g(g(y_{i+1})) = g^2(y_{i+1})$.

Due to the errors inherent in the optical flow method, $y_i = f(y_{i-1}) = g(y_{i+1})$ does not hold universally true, we combine $f(y_{i-1})$ and $g(y_{i+1})$ to approximate y_i. In addition, we introduced two confidence factors ω_f and ω_g to represent the confidence of the optical flow method in the forward and backward directions respectively. Therefore, we use $y_i = \dfrac{\omega_f f(y_{i-1}) + \omega_g g(y_{i+1})}{\omega_f + \omega_g}$ to approximate y_i.

Assuming that subjects only annotate frame $y_{i_1}, y_{i_2}, \cdots, y_{i_n}$, where $i_1 < i_2 < \cdots < i_n$, given reference video x and distorted video $\hat{x}$, we can obtain the entire marking video y by the following equation:

$$y_t = \begin{cases} y_i & \text{if } t \in \{i_1, i_2, \cdots, i_n\} \\ g^{i_1 - t}(y_{i_1}) & \text{if } t < i_1 \\ f^{t - i_n}(y_{i_n}) & \text{if } t > i_n \\ \dfrac{\omega_f^a f^a(y_{i_m}) + \omega_g^b g^b(y_{i_{m+1}})}{\omega_f^a + \omega_g^b} & \text{if } i_m < t < i_{m+1} \end{cases} \tag{1}$$

where $a = t - i_m, b = i_{m+1} - t, m \in \{1, 2, \cdots, n-1\}$. In this way, we can obtain the entire marking video y with only a few annotated frames.

C. Synthetic Data

Due to the time-consuming manual annotation process and the limited diversity in luminance and viewing distance of the manually annotated data, coupled with the substantial need for semantically rich annotations during the training and testing process of the model, we choose to utilize the white-box video visibilty metric FovVideoVDP [19] to generate a large volume of synthetic data for the pre-training stage of the model.

To collect diverse luminance and viewing distance data, we set luminance levels to 10 cd/m^2, 110 cd/m^2, 200 cd/m^2, and configured viewing distance measured in angular resolution as 30, 40, 50, 60 pixel per degree (PPD). Therefore, we built 12 observation environments and applied FovVideoVDP to generate the heatmap across these environments.

However, the heatmaps generated by FovVideoVDP cannot be directly utilized. It is because its pixel values do not represent the probability of an observer perceiving distortion at the corresponding pixel. Instead, they are in just objectionable difference (JOD) units, which stands the level of distortion and do not align with the physical meaning of the manually annotated data. Therefore, we convert the JOD values to probabilities using the following equation:

$$P(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{J(x,y)-\min(J)} e^{\frac{-t^2}{2\sigma^2}} \, \mathrm{d}t \tag{2}$$

where $P(x,y)$ is the probability of an observer perceiving distortion at pixel (x,y), $J(x,y)$ is the JOD value of the pixel, $\min(J)$ is the minimum JOD value in the heatmap, and σ is the standard deviation of the JOD units.

IV. Metric Architecture

A. Display Model

The accuracy and effectiveness of Video Visible Difference Prediction (VDP) relies on accurately simulating the human visual system (HVS), where three core aspects are of most significance, namely, pixel per degree (PPD), brightness (Luminance) and frame rate (FPS) [45].

From such perspective, the display model serves as a preprocess to fuse the luminance, angle resolution, and frame rate conditions into the neural network.

Pixel per degree (PPD) reflects the impact of image resolution on visual clarity. High PPD means higher image details, which are closer to the perception of the human eye in a natural viewing environment. Therefore, this physiological characteristic must be considered when VDP is simulated and evaluated to avoid unnecessary investment of resources beyond the range of human visual resolution [45]. Let PPD denote the Pixels Per Degree, a crucial metric in screen resolution analysis that quantifies the number of pixels visible per degree of visual angle. Its computation is determined by the equation 3:

$$PPD = \frac{R}{W\theta} \tag{3}$$

where R refers to the screen resolution, W represents the screen width, and θ corresponds to the field of view, an angular measure representing the angle subtended by the observer and the screen, which can be further defined by the viewing distance D:

$$\theta = 2 \cdot \arctan\left(\frac{W}{2D}\right) \tag{4}$$

Luminance is another key factor in the display model, which directly affects the perception and visual comfort of the image. The visibility of difference in human eyes can be dramatically distinct by different brightness levels, and the grasp of this dynamic range is crucial in VDP [18]. To account for the variation of luminance, we adopt a transform model by the following equation:

$$L' = L_{\text{black}} + (L_{\text{peak}} - L_{\text{black}}) \ \ln \frac{I - I_{\min}}{I_{\max} - I_{\min}} \tag{5}$$

where I is the original pixel value by R-G-B channels, $I_{\max}$ is the maximal pixel value in the corresponding channel, $I_{\min}$ is the minimal pixel value in the corresponding channel, L_{pear} is the peak luminance, and L_{black} is the luminance when displaying pure black.

Frame rate (frame per second, FPS) is also important in VDP, especially when evaluating video content. FPS affects the smoothness and visual continuity of video streams, while HVS has significant differences in perception of different frame rates, especially when dealing with fast-moving scenes [40]. In order to allow the model to respond to videos with different FPS, we use interpolated frames to align the FPS to the target. The formula is as follows:

$$n = \left\lfloor \frac{\text{FPS}_{\text{target}}}{\text{FPS}_{\text{current}}} + 0.5 \right\rfloor \tag{6}$$

where n is the number of repetitions per frame during frame interpolation, $\text{FPS}_{\text{current}}$ is the frame rate of the video, and $\text{FPS}_{\text{target}}$ is the target frame rate after frame interpolation.

B. UVit Architecture

In this section, we present the architecture of our proposed model, Unified Video and Image Transformer (UViT). UViT is designed to address the task of jointly processing video and image data, leveraging the Transformer architecture to capture both spatial and temporal dependencies effectively.

Let $X_{\text{ref}} \in \mathcal{R}^{B\times C\times F\times H\times W}$ denote the reference image sequence, where B represents the batch size, C denotes the number of channels, F is the number of frames, and $H \times W$ are the spatial dimensions. Similarly, X_{dis} represents the distorted image sequence. Additionally, we incorporate auxiliary information including pixel difference (PPD) and frames per second (FPS).

We employ patch embedding to convert input images into sequences of fixed-size patches. The reference and distorted

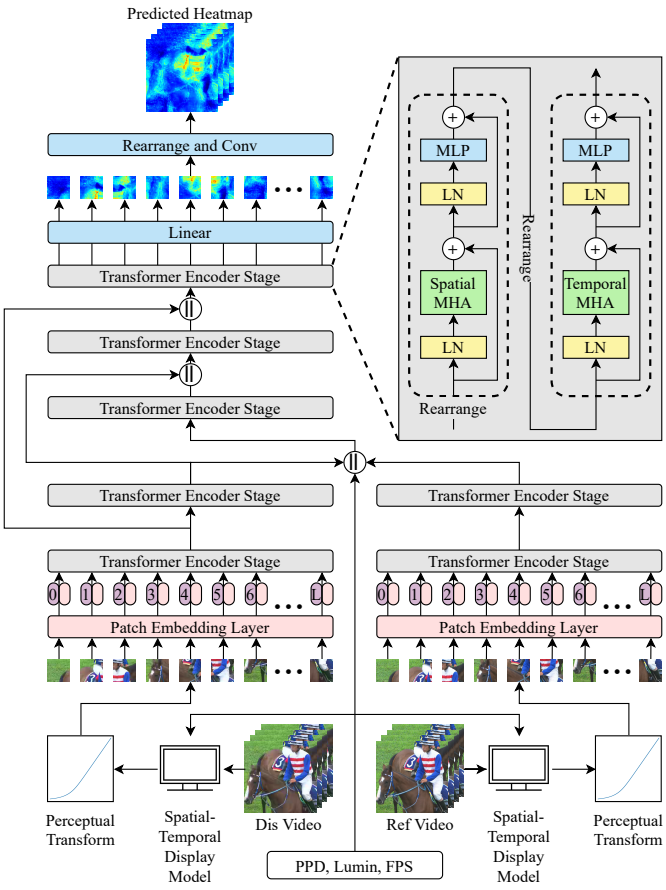

Fig. 5.

image sequences are separately passed through patch embedding layers to obtain patch embeddings:

$$\text{PE}_{\text{ref}} = \text{PatchEmbed}(X_{\text{ref}}), \quad \text{PE}_{\text{dis}} = \text{PatchEmbed}(X_{\text{dis}})$$

Positional embeddings PE_{pos} are added to patch embeddings to provide spatial information. Similarly, temporal embeddings PE_{temp} capture temporal dynamics within the sequences.

$$\text{PE}_{\text{ref}} = \text{PE}_{\text{ref}} + \text{PE}_{\text{pos}}, \quad \text{PE}_{\text{dis}} = \text{PE}_{\text{dis}} + \text{PE}_{\text{pos}}$$

$$\text{PE}_{\text{ref}} = \text{PE}_{\text{ref}} + \text{PE}_{\text{temp}}, \quad \text{PE}_{\text{dis}} = \text{PE}_{\text{dis}} + \text{PE}_{\text{temp}}$$

Both reference and distorted patch embeddings are passed through multiple Transformer encoder layers to capture spatial and temporal dependencies:

$$\text{Out}_{\text{ref}} = \text{Transformer}(\text{PE}_{\text{ref}}), \quad \text{Out}_{\text{dis}} = \text{Transformer}(\text{PE}_{\text{dis}})$$

Then it comes to the middle stage, where the outputs from Transformer encoder layers are concatenated with PPD and FPS information and passed through additional Transformer layers to fuse spatial and temporal features:

$$\text{Out}_{\text{middle}} = \text{Transformer}([\text{Out}_{\text{ref}}, \text{Out}_{\text{dis}}, \text{PPD}, \text{FPS}])$$

The features from the middle stage are combined with skip connections from distorted image encoder layers and processed through the final set of Transformer layers:

$$\text{Out}_{\text{last}} = \text{Transformer}([\text{Out}_{\text{middle}}, \text{Skip}_{\text{dis}}])$$

The output from the last stage is decoded to reconstruct the final image sequence:

$$\hat{X}_{\text{out}} = \text{Decode}(\text{Out}_{\text{last}})$$

where $\hat{X}_{\text{out}}$ represents the reconstructed image sequence.

Finally, the reconstructed image sequence is passed through a convolutional layer to produce the final output.

$$Y = \text{Conv3D}(\hat{X}_{\text{out}})$$

UViT integrates spatial and temporal information using a Transformer architecture, enabling effective processing of both video and image data. Through the proposed architecture, UViT demonstrates promising capabilities in tasks such as image and video restoration and enhancement.

V. Experiment

A. Experimental Setup

For training Deep Video Visible Difference Metric (DV2DM), we employed the probability loss function proposed by [44]. The loss function offers a crucial method to model the experiment data, where it conceptualizes the human annotation process as a sequence of discovery, attention and detection activities to capture the uncertainty in the human-annotated dataset.

Following the pre-processing stage, we split the video into numerous non-overlapping patches and removed the patches where there is no difference between reference patch and distortion patch. Each patch consists of 12 frames and the size of each frame is 48x48.

We implemented DV2DM on the PyTorch v2.0 framework [26] and utilized MMEngine v0.9.0 [8] to perform distributed training on 8 NVIDIA Ampere A100 GPUs. During the training process, we used the adaptive moment estimation with weight decay (AdamW) [15] optimizer with a learning rate of 1×10^{-4}. We partitioned the training process into two stages, as explained below.

B. Training Details

1) Pre-training Stage: Due to the limited variations in luminance, viewing distance and FPS in the manually annotated dataset, we initially conducted our training process on the synthetic dataset containing over 1 million patches. The generation of the synthetic dataset is described in Section III-C. Although the labels generated by the white-box visibilty metric FovVideoVDP [19] may not be entirely accurate, they characterize the general relationship between the original data and labels under different observation environments, aiding the model in capturing these relationships more effectively, which may be lacking in the manually annotated dataset.

2) Fine-tuning Stage: In this stage, we initialize the model's weights with pre-trained weights and fine-tune it using manually annotated data. The data processing workflow during fine-tuning stage remains consistent with the pre-training stage.

C. Result

To evaluate the performance of our model and minimize the bias introduced by training-test split, we split the annotated data into 5 folds and ensured that each data point belongs to only one fold to perform 5-fold cross-validation. We reported the mean and standard deviation of the likelihood used for the loss function after applying the negative log transformation. A higher likelihood corresponds to a smaller negative log-likelihood, indicating a higher accuracy.

We compared DV2DM with other methods and the manually annotated data. The compared methods range from traditional metrics such as MSE, PSNR and SSIM to metrics employing perceptually uniform encoding such as PUMSE, PUPSNR and PUSSIM [46]. In addition, our comparsion includes the deep image VDP, DPVM and the unique but training-free video VDP, FovVideoVDP. Since the manually annotated data includes three difference viewing distances and three different screen luminances, the result of cross-validataion is shown separately in Table II. According to the result, the negative-log-likelihood of DV2DM is significantly lower than FovVideoVDP for all the subsets, demonstrating DV2DM outperforms FovVideoVDP.

D. Ablation Study

To examine whether DV2DM can account for the change of view conditions such as viewing distance, luminance and frames per second, we conduct a series of ablation experiments and report the results in Figure 6. We vary the view conditions for the manual annotation process on identical video contents.

From Figure 6, both FovVideoVDP and DV2DM provide corresponding feedback for different viewing distances, luminance, and FPS, where the general trends are coordinated, while DV2DM yields explicitly better predictions. Specifically, the label points are denser when the PPD is lower or when the luminance is higher, which suggests the common sense that differences can be more visible when either the view distance is closer or the screen is brighter. As for the frame rate, the predicted visible difference is more concentrated on the moving objects than the background with higher FPS and vice versa. This is expected because high frame rates provide less single-image information and more noticeable motion blur, making small changes and details that happen quickly more clearly visible [29]. Though these regular trends reflect the agreement between both predictions of FovVideoVDP and DV2DM and basic HVS experience, it is worth noting that our DV2DM can better predict the differences corresponding to the human vision systems. For instances in Figure 6, the predictions by FovVideoVDP basically mistake by over-estimating the difference in backgrounds.

TABLE II
CROSS-VALIDATION RESULT.

	All	P30L110	P30L200	P30L300	P50L110	P50L200	P50L300	P60L110	P60L200	P60L300
MSE	2.45 ± 1.08	1.78 ± 0.57	2.39 ± 0.71	2.54 ± 0.66	2.51 ± 0.99	2.79 ± 1.10	3.27 ± 1.44	$\mathbf{1.71 \pm 0.78}$	1.84 ± 0.71	2.21 ± 1.05
PSNR	6.70 ± 4.28	7.49 ± 4.17	7.34 ± 4.18	7.30 ± 4.20	6.52 ± 4.35	6.30 ± 4.30	6.36 ± 4.39	6.79 ± 4.19	6.75 ± 4.20	6.64 ± 4.25
SSIM	1.84 ± 0.76	1.46 ± 0.46	1.57 ± 0.39	1.60 ± 0.37	1.94 ± 0.73	1.98 ± 0.86	2.01 ± 0.70	1.86 ± 0.84	1.88 ± 0.84	1.88 ± 0.78
PUMSE	2.69 ± 1.20	1.97 ± 0.67	2.58 ± 0.70	2.75 ± 0.64	2.75 ± 1.14	3.08 ± 1.26	3.56 ± 1.66	1.87 ± 0.81	2.01 ± 0.71	2.42 ± 1.12
PUPSNR	4.88 ± 4.72	5.57 ± 4.65	5.49 ± 4.66	5.45 ± 4.67	4.69 ± 4.75	4.64 ± 4.68	4.64 ± 4.76	4.78 ± 4.71	4.71 ± 4.73	4.65 ± 4.76
PUSSIM	3.46 ± 1.19	2.88 ± 1.08	2.86 ± 0.93	2.88 ± 0.88	3.42 ± 0.98	3.45 ± 0.90	3.35 ± 0.89	4.16 ± 1.54	4.11 ± 1.52	4.03 ± 1.47
DPVM	5.67 ± 3.81	5.78 ± 3.91	6.84 ± 3.61	7.21 ± 3.31	4.89 ± 3.99	5.58 ± 3.70	6.02 ± 3.49	4.63 ± 4.03	5.02 ± 3.88	5.31 ± 3.77
FovVideoVDP	4.73 ± 2.46	5.38 ± 2.84	5.59 ± 2.81	5.74 ± 2.90	4.18 ± 2.23	4.15 ± 2.00	4.59 ± 2.48	4.57 ± 2.17	4.94 ± 2.45	5.10 ± 2.56
DV2DM	$\mathbf{1.72 \pm 0.57}$	$\mathbf{1.29 \pm 0.20}$	$\mathbf{1.35 \pm 0.23}$	$\mathbf{1.43 \pm 0.23}$	$\mathbf{1.59 \pm 0.47}$	$\mathbf{1.96 \pm 0.75}$	$\mathbf{1.72 \pm 0.37}$	1.75 ± 0.45	$\mathbf{1.82 \pm 0.43}$	$\mathbf{1.84 \pm 0.36}$

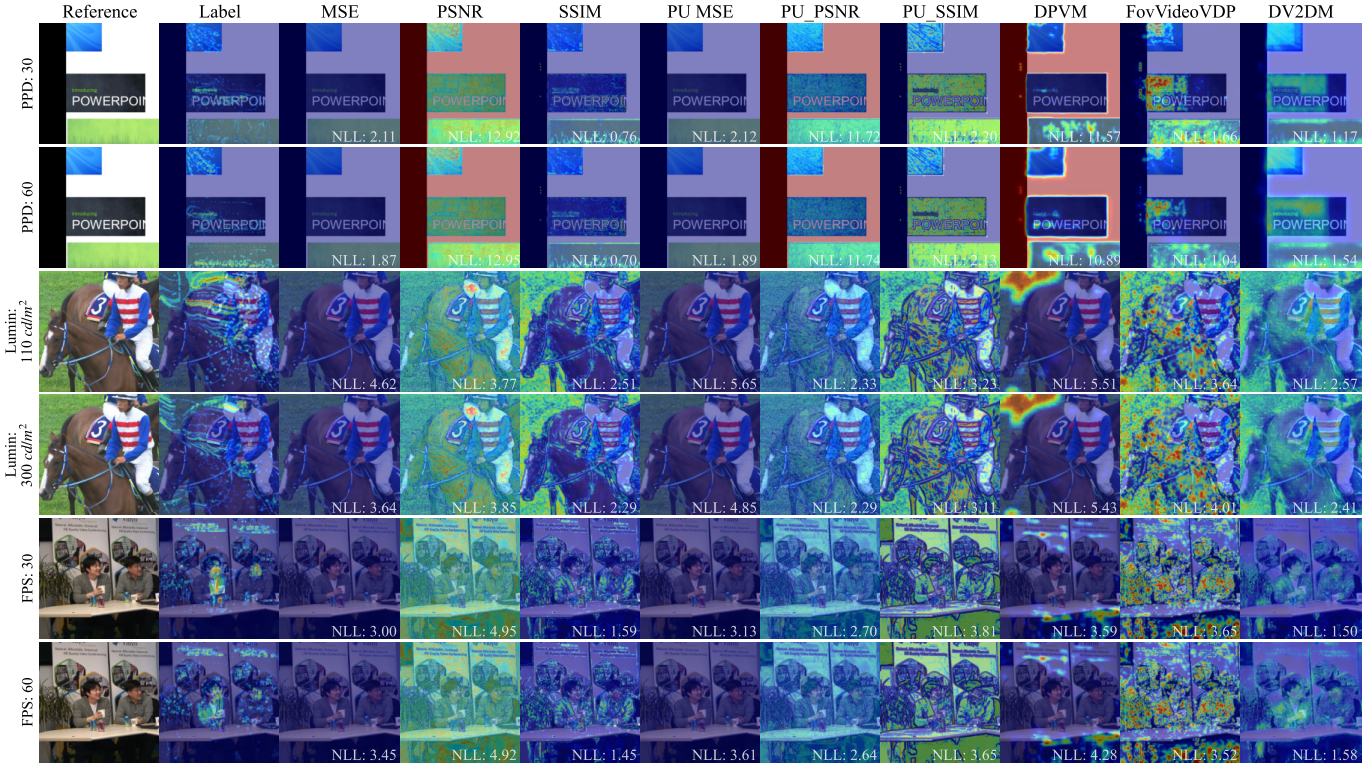

Fig. 6. Reference images, labels and metrics' predictions examples from the manually annotated data. NLL represents negative-log-likelihood.

VI. APPLICATIONS

A. Content-adaptive Watermarking

With the advance in the Artificial Intelligence Generated Content (AIGC) technology, numerous images and videos are used to train AIGC algorithms without authorization from the authors. To protect the copyright of the authors, embedding watermarks stands as a pivotal method. In the case of images or videos, a watermark is often a logo or text that is almost imperceptible but nevertheless real. Hence, the foundation of our application lies in utilizing DV2DM to detect the correct regions for watermark embedding and make the watermark undetectable to the human eye.

In order to embed a watermark, we add a grid of watermark patches in 64x64 pixels to the first frame of the original video. Determining the watermark's intensity involves employing a grid search method. We started at a high watermark intensity so that the difference is clearly visible. Then we iteratively reduce the intensity until DV2DM indicates that the maximum value of all patches for the first frame falls below a predefined threshold. This process is repeated across all the frames to generate the watermarked video.

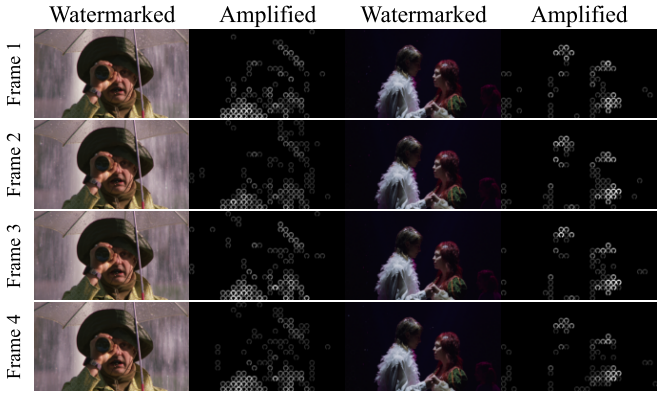

Fig. 7. Examples of content-adaptive watermark application.

Although DV2DM has not been trained specifically for this distortion, the watermark added under the guidance of

DV2DM remains imperceptible, as shown in the Figure 7. The results not only demonstrated DV2DM can detect the suitable regions for watermark addition and guide the watermarking process, but also highlight its adaptability to distortion types absent in the training set. Furthermore, we provided the results of watermarks multiplied by a constant to offer a clearer view of the variations in watermark intensities. For example, the clothing region, where there is a more noticeable color variation and higher frequency, leading to a higher watermark intensity. The intensity remains small in the low-frequency regions.

B. Visually Lossless Video Compression

During the video compression process, inappropriate compression parameters can lead to excessively large file sizes or poor video quality. To balance the trade-off between file size and video quality, visually lossless video compression minimizes file sizes by making the distortion caused by compression practically imperceptible to the human eye. Achieving visually lossless video compression can depend on using appropriate compression parameters, although determining whether a video is visually lossless can pose a challenge.

Hence, we employed HEVC as our video codec, Constant Rate Factor (CRF) as the compression parameter to optimize. and used DV2DM to determine if the compressed video is visually lossless. Beginning with the highest CRF, which renders the video quality unacceptable. We fed the compressed video into DV2DM to generate the distortion heatmap. By averaging the heatmap for each frame, if it is larger than the predefined threshold, we decrease the CRF, improving video quality while increasing file size. This iterative process continues until the averaged heatmap value falls below the threshold. The frames pairs are shown as Figure 8.

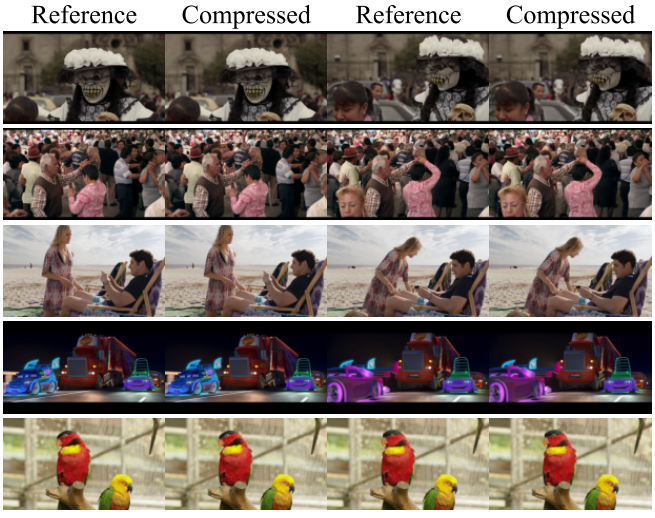

Fig. 8. The frames pairs sampled from the original videos and visually lossless compressed videos.

We set the threshold to 0.01, namely 99% of observers are unable to detect the differences between the compressed videos and the original videos. After obtaining the compressed videos, we recruited 15 volunteers to conduct a Two-Alternative Forced-Choice (2AFC) experiment, where volunteers were asked if they could discern any differences between the compressed and original videos. Based on the 2AFC experiment results, 95% of participants could not perceive any distortion between the two videos. Although 95% is less than 99%, our DV2DM demonstrates some overestimation. Nevertheless, the results highlight the potential of DV2DM guiding the visually lossless video compression.

C. Invisible Adversarial Attacks

In the adversarial attack task, it is necessary to apply imperceptible noise to the input data to induce errors in the predictions of the model. However, a too low noise intensity might fail to allow the model to produce erroneous results while an intense noise may be detected by the human. Therefore, we need to maximize the noise intensity while it is almost imperceptible to the human eye, thereby achieving the invisible undetectable adversarial attack.

Thus, we utilized our DV2DM to detect whether humans can perceive the noise in the task of adversarial attack on the video models. We used Multi-scale Vision Transformers (MViT) [51], [52] as the target video classification model and employed Fast Gradient Sign Method (FGSM) [50] as the adversarial attack approach. We started with a intense noise and iteratively reduce the noise intensity until the maximum value of the heatmap predicted by DV2DM is less than the threshold. The frames sampled from the input video and the corresponding predictions of the model before and after the adversarial attack and are shown as Figure 9.

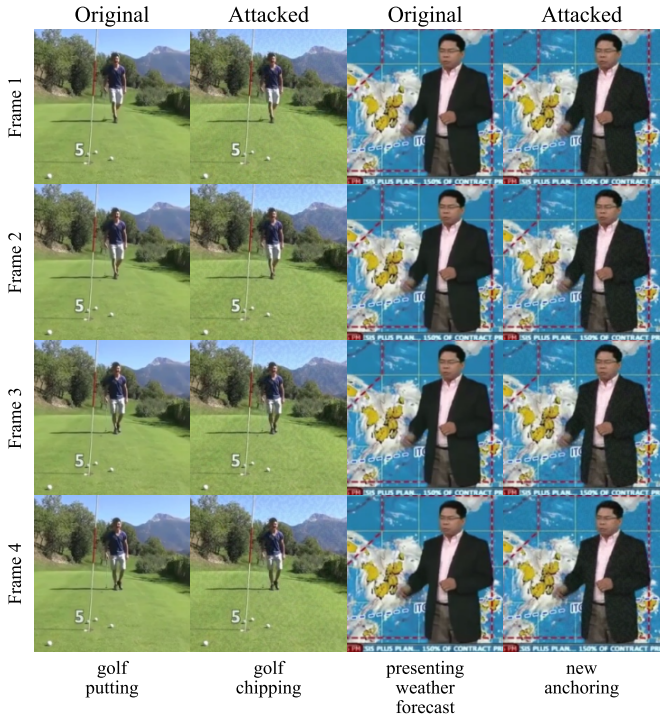

Fig. 9. Frame samples and predictions of the model before and after the adversarial attack.

From the Figure 9, it is clear that there are errors in the predictions of the model, while we cannot perceive the noise in the image after being adversarial attacked. The results indicated that utilizing DV2DM to predict whether the noise is perceivable in the process of undetectable adversarial attack is effective.

D. Video Super-Resolution Metric

In the field of video research, the video super-resolution is also an essential task. It aims to convert low-resolution video to high-resolution while minimizing the distortion caused by the super-resolution algoithms. To evaluate the video quality after super-resolution, Mean Squared Error (MSE) and Structural Similarity (SSIM) are often used as the metrics. However, both MSE and SSIM are which are evaluated at the pixel level which is different from the perception process of the human eye. Thus MSE and SSIM are difficult to exactly and accurately describe perception quality. Therefore, we adopted DV2DM as an evaluation metric for the video super-resolution task to assess the distortion introduced by the super-resolution process.

We downsampled the video with a resolution of 3840x2160 by ffmpeg before restoring it to the original resolution by an interpolation algorithm. Then we predicted the distortion between these two videos using MSE, SSIM and DV2DM respectively and get the predicted heatmap. The results are shown as Figure 10.

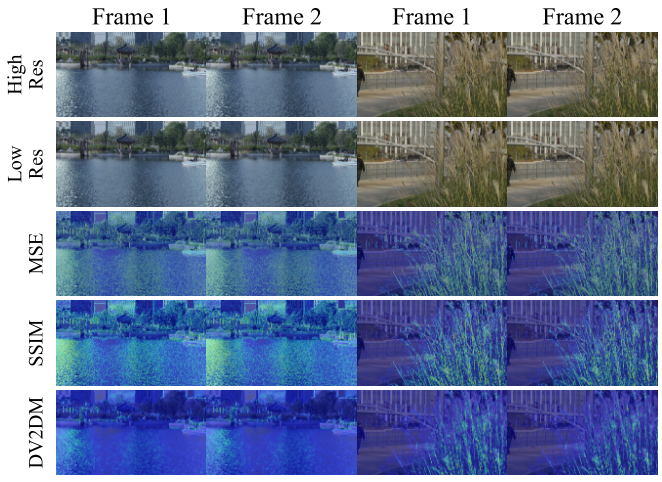

Fig. 10. The heatmaps predicted by Mean Squared Error (MSE), Structural Similarity (SSIM) and DV2DM based on high-resolution frames and low-resolution frames after super-resolution.

From the Figure 10, it can be seen that MSE, SSIM and DV2DM keep the consistency in the prediction trend, which shows the reasonableness of DV2DM as an evaluation metric for the video super-resolution task. Meanwhile, it can be noticed that the heatmap predicted by DV2DM is smoother compared with MSE and SSIM, which is more in line with the human eye's perception of video quality. Therefore, DV2DM can be used as an evaluation metric for the video super-resolution task to simulate the human eye's perception of the video after super-resolution.

VII. Conclusion

VDP within the realm of video processing remains an underexplored domain, presenting multiple challenges, including a shortage of datasets and the distinct temporal and spatial attributes of video data. This research endeavors to investigate the applicability of VDP to video content, along with examining the influence of various viewing conditions such as luminance, angular resolution, and playback rate, which are crucial considerations. We introduce a two-stage model, the Deep Video Visible Difference Metric (DV2DM), and establish a comprehensive dataset, dubbed ViLocVis, to encapsulate the nuances of perceptible differences as discerned by the human visual system. Our empirical findings demonstrate that the ViLocVis dataset accurately embodies the anticipated VDP experience, while the DV2DM proficiently captures the essence of video VDP, reliably forecasting perceptible disparities across diverse viewing scenarios. Ultimately, this study addresses the void in video-based VDP research, offering both a robust dataset and a methodological framework that serve as valuable references for future explorations in this burgeoning field.

References

[1] Ahn, S., Choi, Y., and Yoon, K. Deep learning-based distortion sensitivity prediction for full-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 344–353, 2021.

[2] Akyazi, P. and Ebrahimi, T. Learning-based image compression using convolutional autoencoder and wavelet decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, number CONF, 2019.

[3] Aydin, T. O., Mantiuk, R., Myszkowski, K., and Seidel, H.-P. Dynamic range independent image quality assessment. *ACM Trans. Graph.*, 27(3): 1–10, aug 2008. ISSN 0730-0301. doi: 10.1145/1360612.1360668. URL https://doi.org/10.1145/1360612.1360668.

[4] Banterle, F., Artusi, A., Moreo, A., and Carrara, F. Nor-vdpnet++: Efficient training and architecture for deep no-reference image quality metrics. In *ACM SIGGRAPH 2021 Talks*, pp. 1–2. 2021.

[5] Charrier, C., Lebrun, G., and Lezoray, O. A machine learning-based color image quality metric. In *CGIV*, pp. 251–256, 2006.

[6] Chen, C., Shi, Y. Q., and Su, W. A machine learning based scheme for double jpeg compression detection. In *2008 19th international conference on pattern recognition*, pp. 1–4. IEEE, 2008.

[7] Cheon, M. and Lee, J.-S. Subjective and objective quality assessment of compressed 4k uhd videos for immersive experience. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(7):1467–1480, 2018. doi: 10.1109/TCSVT.2017.2683504.

[8] Contributors, M. MMEngine: Openmmlab foundational library for training deep learning models. 2022.

[9] Daly, S. J. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, volume 1666, pp. 2–15. SPIE, 1992.

[10] Illumination, I. Recommendations on uniform color spaces. *Color-Difference Equations, Psychometric Color Terms, CIE*, 1978.

[11] Ji, R., Karam, L. J., et al. Learning-based visual compression. *Foundations and Trends® in Computer Graphics and Vision*, 15(1):1–112, 2023.

[12] Kim, W., Kim, J., Ahn, S., Kim, J., and Lee, S. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 219–234, 2018.

[13] Li, T., Min, X., Zhao, H., Zhai, G., Xu, Y., and Zhang, W. Subjective and objective quality assessment of compressed screen content videos. *IEEE Transactions on Broadcasting*, 67(2):438–449, 2020.

[14] Liu, W., Duanmu, Z., and Wang, Z. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM Multimedia*, pp. 546–554, 2018.

[15] Loshchilov, I. and Hutter, F. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

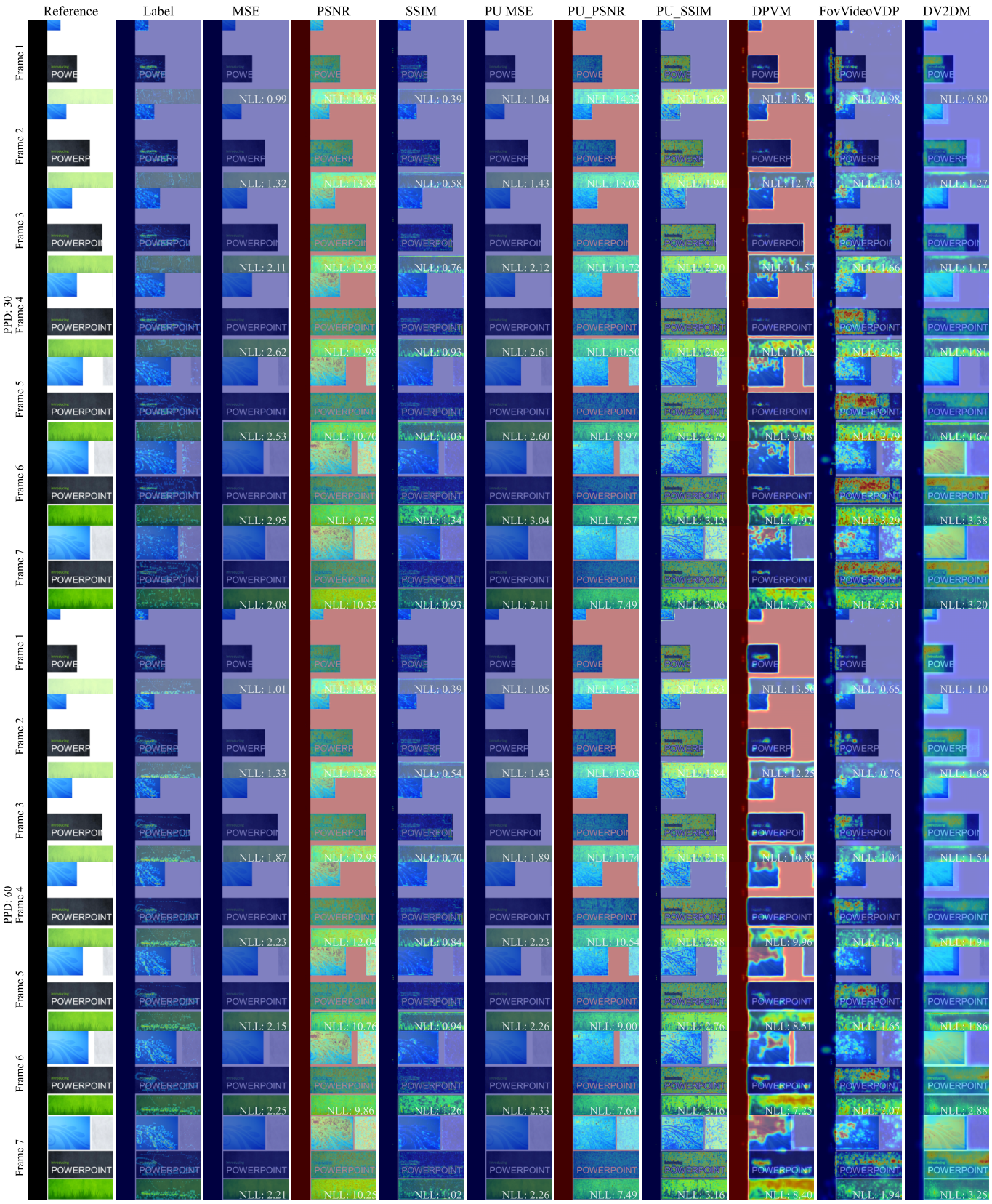

Fig. 11. Reference images, labels and metrics' predictions examples under different PPDs.

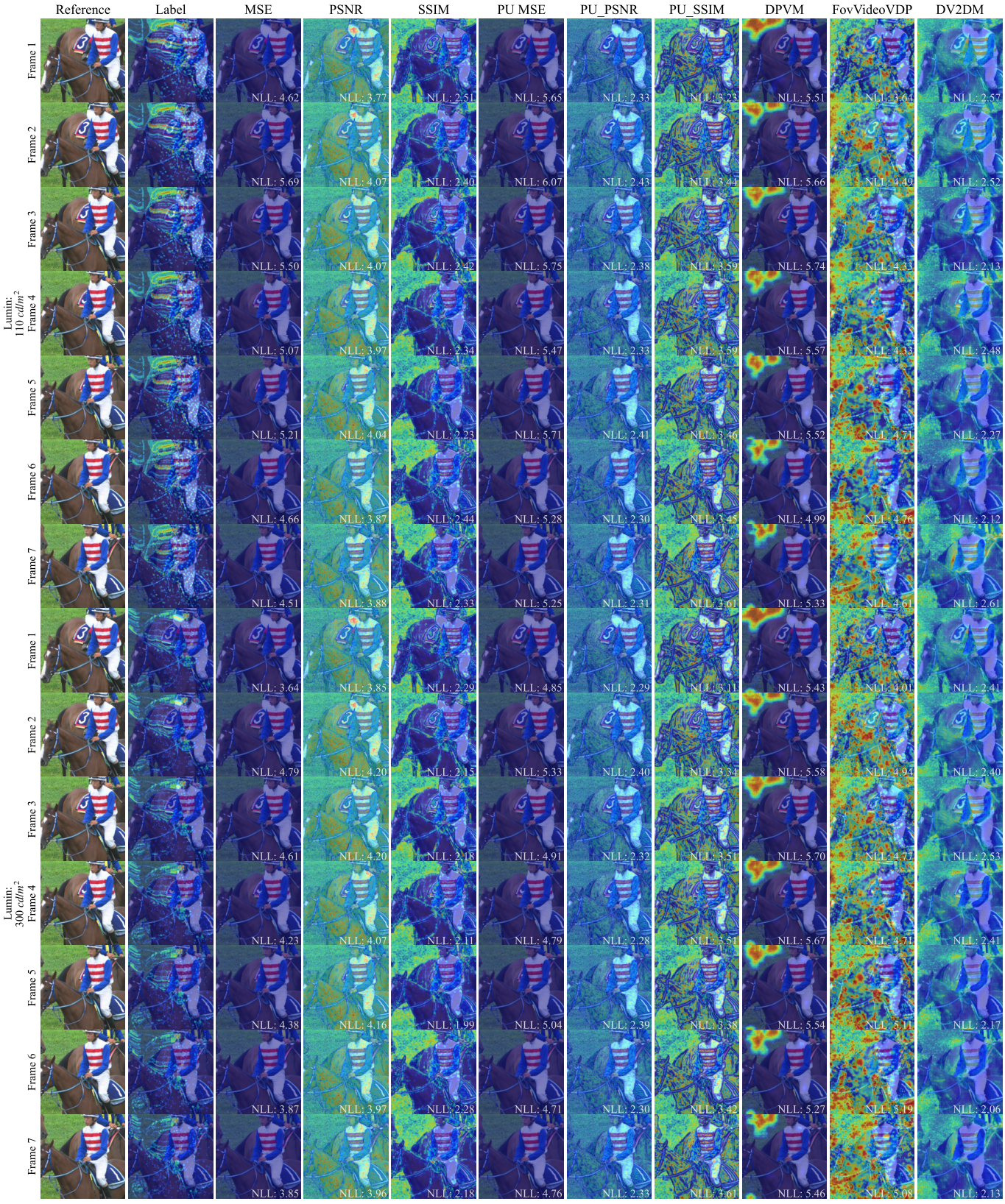

Fig. 12. Reference images, labels and metrics' predictions examples under different luminances.

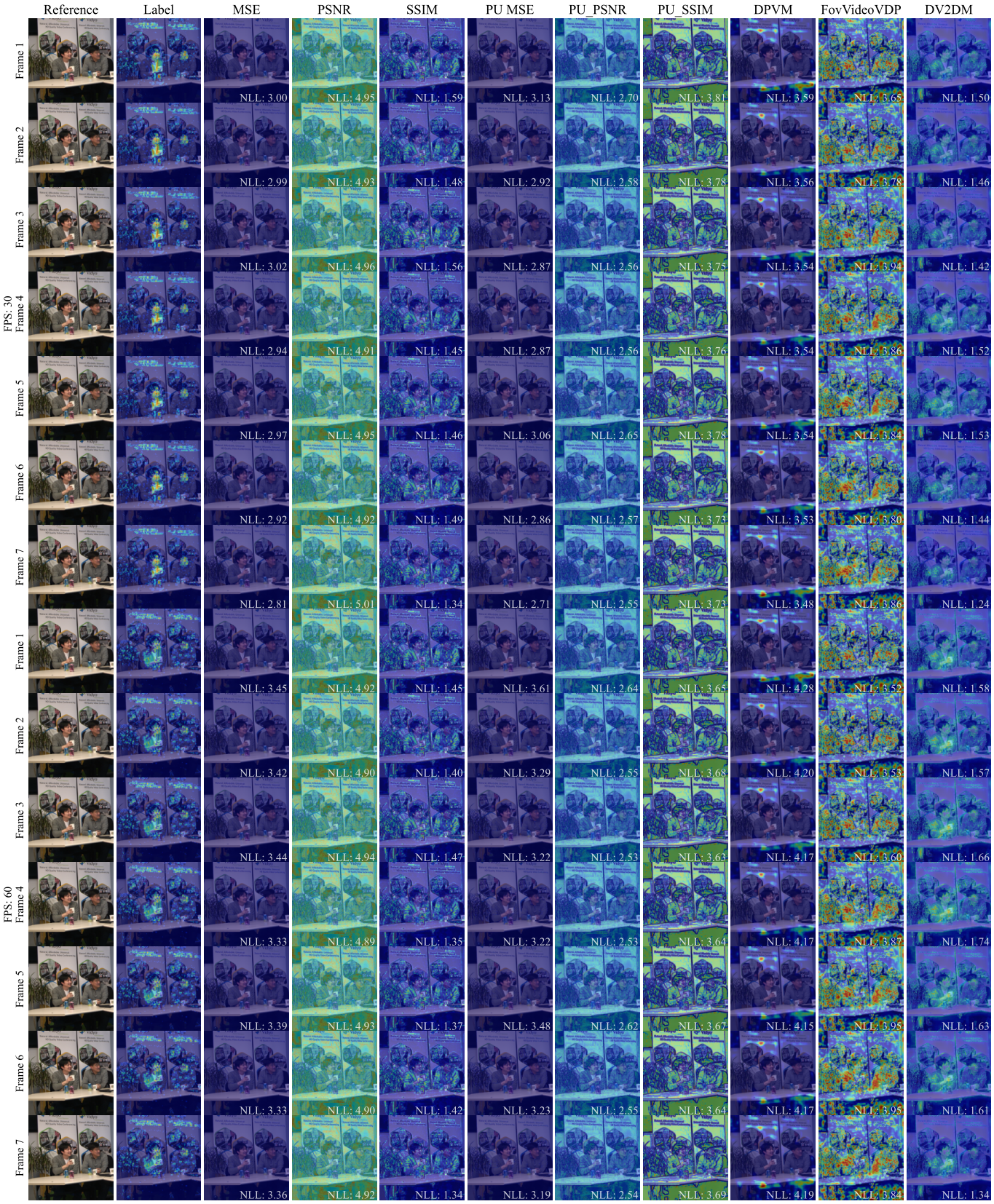

Fig. 13. Reference images, labels and metrics' predictions examples with different FPSs.

[16] Lu, G., Zhong, T., Geng, J., Hu, Q., and Xu, D. Learning based multi-modality image and video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6083–6092, 2022.

[17] Mantiuk, R., Daly, S. J., Myszkowski, K., and Seidel, H.-P. Predicting visible differences in high dynamic range images: model and its calibration. In *Human Vision and Electronic Imaging X*, volume 5666, pp. 204–214. SPIE, 2005.

[18] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

[19] Mantiuk, R. K., Denes, G., Chapiro, A., Kaplanyan, A., Rufo, G., Bachy, R., Lian, T., and Patney, A. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40 (4):1–19, 2021.

[20] Mercat, A., Viitanen, M., and Vanne, J. Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20, pp. 297–302, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368452. doi: 10.1145/3339825.3394937. URL https://doi.org/10.1145/3339825.3394937.

[21] Min, X., Zhai, G., Zhou, J., Farias, M. C. Q., and Bovik, A. C. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29:6054–6068, 2020. doi: 10.1109/TIP.2020.2988148.

[22] Nadenau, M. J. and Reichel, J. Compression of color images with wavelets considering the hvs. In *Human Vision and Electronic Imaging IV*, volume 3644, pp. 129–140. SPIE, 1999.

[23] Narwaria, M. and Lin, W. Svd-based quality metric for image and video using machine learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):347–364, 2011.

[24] Narwaria, M., Mantiuk, R. K., Da Silva, M. P., and Le Callet, P. Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24 (1):010501–010501, 2015.

[25] Palmer, S. E. *Vision science: Photons to phenomenology*. MIT press, 1999.

[26] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

[27] Peli, E. *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*, volume 2. World scientific, 1995.

[28] Pöppel, E. A hierarchical model of temporal perception. *Trends in cognitive sciences*, 1(2):56–61, 1997.

[29] Selfridge, R., Noland, K. C., and Hansard, M. Visibility of motion blur and strobing artefacts in video at 100 frames per second. In *Proceedings of the 13th European Conference on Visual Media Production (CVMP 2016)*, pp. 1–10, 2016.

[30] Shahid, M., Rossholm, A., and Lövström, B. A no-reference machine learning based video quality predictor. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 176–181. IEEE, 2013.

[31] Shahid, M., Rossholm, A., Lövström, B., and Zepernick, H.-J. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP Journal on image and Video Processing*, 2014:1–32, 2014.

[32] Shang, Z., Ebenezer, J. P., Bovik, A. C., Wu, Y., Wei, H., and Sethuraman, S. Assessment of subjective and objective quality of live streaming sports videos. In *2021 Picture Coding Symposium (PCS)*, pp. 1–5, 2021. doi: 10.1109/PCS50896.2021.9477502.

[33] Shang, Z., Ebenezer, J. P., Wu, Y., Wei, H., Sethuraman, S., and Bovik, A. C. Study of the subjective and objective quality of high motion live streaming videos. *IEEE Transactions on Image Processing*, 31:1027–1041, 2022. doi: 10.1109/TIP.2021.3136723.

[34] Song, L., Tang, X., Zhang, W., Yang, X., and Xia, P. The sjtu 4k video sequence dataset. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 34–35, 2013. doi: 10.1109/QoMEX.2013.6603201.

[35] Sreelekha, G. and Sathidevi, P. An hvs based adaptive quantization scheme for the compression of color images. *Digital Signal Processing*, 20(4):1129–1149, 2010.

[36] Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. doi: 10.1109/TCSVT.2012.2221191.

[37] Wang, H., Gan, W., Hu, S., Lin, J. Y., Jin, L., Song, L., Wang, P., Katsavounidis, I., Aaron, A., and Kuo, C.-C. J. Mcl-jcv: A jnd-based h.264/avc video quality assessment dataset. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1509–1513, 2016. doi: 10.1109/ICIP.2016.7532610.

[38] Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

[39] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[40] Watson, A. B. High frame rates and human vision: A view through the window of visibility. *SMPTE Motion Imaging Journal*, 122(2):18–32, 2013.

[41] Watson, A. B., Hu, J., and McGowan, J. F. Digital video quality metric based on human vision. *Journal of Electronic imaging*, 10(1):20–29, 2001.

[42] Winkler, S. Perceptual video quality metrics—a review. *Digital video image quality and perceptual coding*, pp. 155–180, 2017.

[43] Winkler, S. and Mohandas, P. The evolution of video quality measurement: From psnr to hybrid metrics. *IEEE transactions on Broadcasting*, 54(3):660–668, 2008.

[44] Wolski, K., Giunchi, D., Ye, N., Didyk, P., Myszkowski, K., Mantiuk, R., Seidel, H.-P., Steed, A., and Mantiuk, R. K. Dataset and metrics for predicting local visible differences. *ACM Transactions on Graphics (TOG)*, 37(5):1–14, 2018.

[45] Ye, N., Wolski, K., and Mantiuk, R. K. Predicting visible image differences under varying display brightness and viewing distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5434–5442, 2019.

[46] R. K. Mantiuk and M. Azimi. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR, *2021 Picture Coding Symposium (PCS)*

[47] Zhang, L., Zhang, L., Mou, X., and Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[48] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[49] Zhang, X., Wandell, B. A., et al. A spatial extension of cielab for digital color image reproduction. In *SID international symposium digest of technical papers*, volume 27, pp. 731–734. Citeseer, 1996.

[50] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[51] Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., and Feichtenhofer, C. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

[52] Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., and Feichtenhofer, C. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804–4814, 2022.