***Machine Learning in Video Editing***

1. **Assets Collection**

Collecting the necessary data for the trials was one of the primary challenges of this paper, as it is for other contemporary neural network models. Even though there are multiple datasets for our distinct problems - video super resolution, denoising, and enhancement - making one that incorporates all of these remains challenging. Fortunately, having enough computational power, a dataset in this manner can be created by taking shots of movies before and after edit, while also reducing their resolution. In this manner, for VSR we used BSDS500 dataset for training, BSDS100 for validation and Set5 for testing. By using a SRResNet model, we were able to cover the denoising problem, especially by preparing two datasets that were used in others papers, REDS and VIMEO90k. We utilized the Rendered WB dataset for video improvement, which comprises two sets. The first set only included the initial portion for the tests, yielding about 62k input photos.

Since the video we would get as input would be divided into frames, each of which would be altered using both models to improve resolution and color, we utilized datasets for the tests that actually contain distinct pictures rather than continuous images. For future work, a dataset which contains edited and non-edited continuous images or videos could help in creating a connection between each frame of a shot. In addition, the shot classification dataset [1] may contribute a lot in how to edit a video, for example whether it’s indoor or outdoor, you can choose easier how to enhance the color of the images. The single problem in this approach is the computational power, which cannot be done on a single laptop or computer, especially if you choose an unsupervised learning method, such as GAN.

1. **Experiments**
   1. **Dataset**

We use a train-validation-test split for the first problem, and only a train-test split for the second one. With regard to the VSR, we use a 500-100-5 picture partition for experimental purposes. Since these groups of photographs do not share a picture, they are considered distinct sets. In case of the REDS dataset, we use blur bicubic images as input, which contains 100 folders with each 100 continuous photos, having a 4x downscale, and the sharp pictures as ground truth. For the enhancement model, we use a 80-20 split while maintaining the creation of disjoints sets for our model, resulting in better evaluation and results.

* 1. **Evaluation metrics**

The metrics for both models are the peak signal-to-noise ratio, which is based on pixel-level comparison, and structural similarity, which compares three aspects of the images: luminance, contrast, and structure. This comparison yields a more accurate measurement while adding mean absolute error for VSR. The loss function for both models is the mean square error.

* 1. **Implementation details**

We use SRResNet [2] and DeepWBNet [3] for our problems which we train using the collected datasets. With the addition of an Adam optimizer, each model was adjusted in accordance with the datasets, yielding better than average outcomes. To process the footage, we split it up into frames using OpenCV, then we go through each model to adjust it before redistributing it back into a video.

* 1. **Following work**

In order to fine-tune this model even further, we propose further investigation in how the classification of scenes can influence the model’s behavior in choosing what enhancement to apply to each frame, especially when having continuous images that can be used in determining which object will be in focus next, or if there will be completely other scene.

* 1. **Experimental result**
     1. **SRResNet**

Our computational capacity limited us to using the BSDS datasets; we were unable to run the model on the REDS, but we were still able to obtain 0.89 structural similarity scores and a peak signal-to-noise ratio of about thirty. This method has drawbacks since it requires the image's output as well for comparison, which makes it challenging to build datasets and produce accurate results. In light of this, applying a GAN to the super-resolution problem would be a more beneficial strategy.

* + 1. **DeepWBNet**

We obtain worse findings as the other network, with major differences, with 0.5 structural similarity and a peak signal-to-noise ratio of roughly 15, a major cause being not being able to trained on a larger dataset. This raises the possibility that, by adding scene classification to the fine-tuning process, we could achieve even better video editing outcomes and generate fresh perspectives on the topic.

1. **Related Work**

A number of studies on super resolution in photos and films have offered many approaches to enhancing this task [2], [4], [5], including training a model to create new pixels or enlarge ones that already exist. Nevertheless, the work “*Activating More Pixels in Image Super-Resolution Transformer*” [6] still remains the finest resource for achieving SR in photos, with the top scores across nearly all 2x and all 4x upscaling datasets, which developed a Hybrid Attention Transformer, that was pre-trained on ImageNet, then trained on the DF2K (Div2K + Flicker2K) dataset.

There are fewer studies on the subject of video color enhancement, albeit they all focus on different aspects of the subject, such as exposure correction [7], colorizing images [8], or white-balance editing [3] [9]. This is still an unexplored topic that has not received much attention, partly because it is difficult to collect the "right" images-which can vary depending on who is viewing them-and partly because there aren't enough datasets accessible. Progress it’s being made, but there still is not something that concludes them together.

As a result, there is currently no study that integrates the two, which makes it exceedingly challenging to approach the topic and develop improvements.

**References**

[1] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson and In So Kweon - The Anatomy of Video Editing: A Dataset and Benchmark Suite for AI-Assisted Video Editing; *Computer Vision – ECCV*, 2022.

[2] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi - Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network; *CVPR.2017.19*, 2017.

[3] Mahmoud Afifi, Michael S. Brown - Deep White-Balance Editing; *CVPR42600.2020.00147*, 2020.

[4] Renjie Liao, Xin Taom Ruiyu Li, Ziyang Ma, Jiaya Jia - Video Super-Resolution via Deep Draft-Ensemble Learning; *IEEE International Conference on Computer Vision*, 2015.

[5] Alice Lucas, Santiago Lopez-Tapia, Rafael Molina and Aggelos K. Katsaggelos - Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution; *IEEE International Conference on Image Processing*, 2018.

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, Chao Dong - Activating More Pixels in Image Super-Resolution Transformer; *CVPR52729.2023.02142*, 2023.

[7] Mahmoud Afifi, Konstantinos G. Derpanis, Bjorn Ommer, Michael S. Brown - Learning Multi-Scale Photo Exposure Correction; *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] Ligeng Zhu, Brian Funt - Colorizing Color Images; *Proc. IS&T Int’l. Symp. on Electronic Imaging: Human Vision and Electronic Imaging*, 2018.

[9] Mahmoud Afif, Brian Price, Scott Cohen, Michael S. Brown - When Color Constancy Goes Wrong: Correcting Improperly White-Balanced Images; *CVPR.2019.00163*, 2019.