Modular Memorability: Tiered Representations for Video Memorability Prediction

Théo Dumont
Mines Paris, PSL – Research University
Paris, France
dumont.theo@protonmail.com

Juan Segundo Hevia
Memorable AI
Boston, USA
juan.hevia@memorable.io

Camilo L. Fosco[*]
Memorable AI
Boston, USA
camilo@memorable.io

Abstract

The question of how to best estimate the memorability of visual content is currently a source of debate in the memorability community. In this paper, we propose to explore how different key properties of images and videos affect their consolidation into memory. We analyze the impact of several features and develop a model that emulates the most important parts of a proposed "pathway to memory": a simple but effective way of representing the different hurdles that new visual content needs to surpass to stay in memory. This framework leads to the construction of our M3-S model, a novel memorability network that processes input videos in a modular fashion. Each module of the network emulates one of the four key steps of the pathway to memory: raw encoding, scene understanding, event understanding and memory consolidation. We find that the different representations learned by our modules are non-trivial and substantially different from each other. Additionally, we observe that certain representations tend to perform better at the task of memorability prediction than others, and we introduce an in-depth ablation study to support our results. Our proposed approach surpasses the state of the art on the two largest video memorability datasets and opens the door to new applications in the field. Our code is available at https://github.com/tekal-ai/modular-memorability.

1. Introduction

The human brain is optimized to remember important content and forget irrelevant information. Research has shown that in the world of visual imagery, the brain's recall ability is influenced by the content itself: certain images and videos tend to stay in memory for longer, no matter the audience it is shown to or the context it appears in [3, 9]. The property of visual content that makes it more or less memorable is referred to as memorability, and current research studies this phenomenon as an intrinsic property. Memorability has been shown to be highly consistent across observers [10, 13, 28, 40], uncorrelated with aesthetics [28, 29] and highly unintuitive [29]. Some studies are proposing that it might be a proxy to the utility of the information carried by visual content, as measured by the human brain [9].

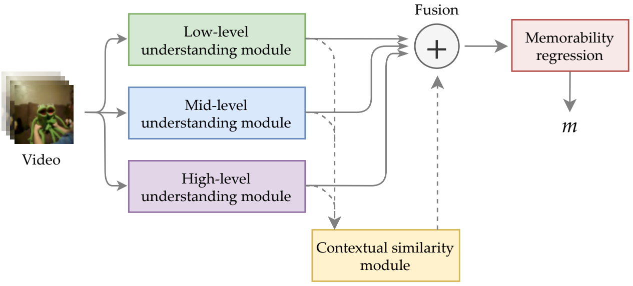

Video
Low-level understanding module
Mid-level understanding module
High-level understanding module
Contextual similarity module
Fusion
+
Memorability regression
m

Figure 1. **Our proposed modular framework.** Our framework predicts memorability by extracting low-level, mid-level and high-level memorability-aware representations. These representations are compared to a predefined visual context to extract features measuring similarity with this given context. Our M3-S model utilizes four modules to obtain these representations: a low-level understanding module composed of traditional feature extractors, a mid-level understanding module focused on scene and object properties, a high-level understanding module that extracts temporal patterns and actions, and a contextual similarity module that computes features through clustering. The feature vectors produced by the modules are fused and fed to a regression module to produce memorability scores.

Given its consistency, many previous works have tried to develop systems to predict memorability scores from visual media directly. Some developments attempt to use low-level image features and specific semantic information [29, 30], while more recent work has focused on deep neural networks, leveraging their ability to learn rich representations through regressing ground-truth scores directly from the pixel-level visual input [31]. Here, we argue that current

[*]Corresponding author.

DNN approaches are trying to solve the problem with black-box predictors that do not leverage the underlying structure governing memorability. Indeed, previous work [42] shows that our brain processes visual stimuli by first aggregating low level patterns (early visual areas V1 and V2), then understanding the contents of the scene (higher visual areas V3A, V4v, V7), and finally integrating the meaning of the event being witnessed and linking with previous knowledge (prefrontal cortex). Although some of the existing systems work on different dimensions of the input (optical flow, raw pixels, text descriptions), they tend to overlook predictive patterns that can be acquired through a specific modeling of low-level, mid-level and high-level representations. Specifically, it has been shown that memorability is sensitive to a set of specific properties [29] (that we define and expand on in this work), such as clutter, camera movement, distinctiveness of objects, and other semantic and cognitive dimensions. Some of these properties are considered low-level: they correspond to simple transformations of the raw pixel input, photometric properties, clarity of image, or properties of the capturing process (blurriness, camera movement, *etc*). Other properties can be considered mid-level, such as the composition of the scene, the type of objects in it, the general setting, *etc*. Finally, high-level properties are usually related to the action depicted, emotion transmitted by the content, or general goals of the actors.

We propose a new memorability framework that explicitly models these three categories by instantiating *modules* that are specifically designed to extract representations that are relevant for memorability, and representative of each category. We call these representations *tiered*, as each representation captures information from a different tier (low, mid and high) of memorability properties. Our modular memorability model additionally introduces a fourth module that computes representations capturing the similarity of a given input with its most likely visual context: modeling this final property is key to understand contextual effects on memorability. To define these modules, we perform an in-depth analysis over the factors that influence memorability. We show that each of these modules contribute to memorability in their own way, that the representations they yield are more interpretable than black box counterparts, and that combining the information from these representations yields competitive models on the two main datasets for video memorability: VideoMem [13] and Memento10k [40].

To summarize, our key contributions are:

1. We introduce a comprehensive analysis of the factors that influence memorability, leading to a categorization in *tiers* that we leverage to propose a new modular framework to learn representations which capture the essence of each tier;
2. We propose a novel memorability model based on these modules, M3-S, that combines information from different tiers, contrasts it with contextual data, and uses it to perform competitively on VideoMem and Memento10k;
3. We perform an in-depth ablation study of the model to obtain key insights about each tier of representations, such as their potential for interpretability, their impact on model performance and the feature representations they learn.

2. Related Work

Image memorability. The field of memorability started with the discovery that intrinsic image properties directly influence recall probabilities on randomized pools of subjects [29]. Importantly, the memorability metric has been shown to be highly consistent across observers [3, 9, 29, 32], which makes it predictable: Khosla *et al*. [31] collected a large scale image memorability dataset (LaMem) and developed neural networks that can predict memorability scores from pixel inputs; Bylinskii *et al*. [10] found that some extrinsic factors such as context can influence memorability, and built models to predict that effect; [41] introduced models that achieve close to human performance on the LaMem dataset. Several other works have studied specific aspects of memorability, such as the memorability of faces [4], objects [18], scenes [36] or specific categories [25]. Although we focus on video memorability prediction, the datasets and concepts in this subarea are relevant to our framework.

Video memorability. Following the success of image memorability, several works have worked on advancing the field of video memorability prediction. Large datasets were recently collected by the community [13, 40], and models were introduced in these works that take into account motion and visual appearance to make their predictions. The MediaEval competition [11] has additionally introduced several new works that explore this field [15, 17, 22, 46]. Others have worked to connect memorability to brain imaging: Han *et al*. [27] proposes video memorability models that use fMRI features to make predictions. Finally, Shekhar *et al*. [43] introduce video memorability models that can estimate sub-shot memorability and perform competitively in video summarization tasks. Importantly, all of these methods utilize black box networks that don't explicitly separate low, mid and high-level visual features, while our model does this by design.

3. Key properties impacting memorability

3.1. Key properties

In this section, we introduce an in-depth analysis and a categorization in tiers of the factors that influence memorability, based on previous work and our own analysis. The low, mid and high-level factors detailed below are purely intrinsic to an image or a video, whereas the contextual factors depend on the external context of the visual content and involve the notion of distinctiveness between videos.

Low-level features. *Color*. Mean hue and saturation have been shown to be weakly correlated to uncorrelated to memorability [28, 29, 32]. However, brightness and contrast — defined as the mean and standard deviation of the value component of HSV — do have a weak positive correlation with memorability [18]. This observation is supported by the work of Goetschalckx *et al*. [23], whose GAN model tend to produce more memorable images by increasing their brightness and colorfulness.
Motion. Newman *et al*. [40] observed that static videos often have a low memorability, and found that leveraging motion information such as optical flow allows to refine the memorability predictions of frame-based models. Basavaraju *et al*. [5, 7] as well as Han *et al*. [27] use explicit motion cues along with input images to enhance the performance of their model. Extreme motion however, for instance when the camera motion is uncontrollably high, can make the video memorability drop drastically, except if a very specific element can be attended or is clearly identifiable in the center of focus.

Mid-level features. *Scene composition*. Previous work has shown that scene composition and complexity encompass some memorability cues. Whereas Han *et al*. [27] used the number of regions or the amount of contours in the video as their model features, Dubey *et al*. [18] showed that the memorability of a scene drops when the number of objects in it goes up, and that this memorability simultaneously becomes more difficult to predict. Goetschalckx *et al*. [24] characterized "good visual organizations" as the being both easily processed and robust against transformation and found that these two metrics correlate moderately with memorability.
Saliency. Multiple studies explored the predictive capacity of saliency over memorability [2, 18, 27, 31, 43]. Khosla *et al*. [31] demonstrated that scene involving a specific point of focus are more easily remembered; Dubey *et al*. [18] showed that the number of fixation counts on an object is robustely correlated with that object's memorability, and that saliency is a good predictor of object memorability in simple contexts with only few objects. The AMNet model [21] leverages these findings and uses a soft attention mechanism to refine its memorability score prediction three times, gaining performance while also confirming through visualizations that images with concentrated regions of interest tend to be more memorable than those whose visual content is spread out across the frame.
Object semantics. Several studies demonstrated the importance of object semantics in the context of visual memorability prediction [18, 25, 28, 29, 32, 41, 44]. In particular, through the study of images whose objects had been segmented and annotated beforehand, Dubey *et al*. [18] observed that "image memorability is greatly affected by the memorability of its most memorable object". Human-related objects, such as faces or body parts, generally account for very memorable scenes, as opposed to landscapes and inanimate items. The essential predictive capacity of object semantics led to their integration to multiple competitive memorability prediction models [21, 39–41, 44].

High-level features. *Actions*. Actions and other high-level scene semantics play an very important role in steering the memorability of a video. Early work pointed this importance out along with that of the object semantics in their analyses [28, 29, 32], and the more recent work of Perera *et al*. [41] showed that scene classification cues are even more predictive than object classification cues regarding memorability — the better option being to use both at the same time. Here also, human-related scenes are way more memorable: actions taking place in interiors have much higher memorability scores than landscapes and natural scenes [29, 41]. Recent video memorability models often involve a feature extractor or a branch that focuses on unmasking the high-level scene semantics of the input videos, for instance using image or video captioning [12, 14, 40, 43].
Emotions. Emotionally salient objects and scenes increase memorability [9, 31], but some emotions are more memorable than others. For instance, images evoking disgust or amusement are statistically more memorable that images showing most other emotions; and overall, negative emotions such as anger and fear tend to be more memorable than those portraying positive ones [9, 31]. This results are supported by the work of Goetschalckx *et al*. [23], where emotion-evoking portraits tend to be more memorable than the others. To predict memorability, several recent work used emotion cues, either categorical [6] or textual [14].
Uncorrelated features. It is also important to note the high-level features that do not correlate well with visual memorability, even though they intuitively seem to. For instance, despite that memorability is highly consistent across observers, people are bad at predicting if an image will be memorable or not. Even worse, human estimation of memorability is negatively correlated with actual memorability [26, 28]. Moreover, the aesthetics of images and their interestingness are not correlated with memorability. Interestingly, aesthetics does correlate well with assumed memorability, as observers tend to have the wrong intuition that beautiful and interesting images will produce a lasting memory [9, 28, 31]. Whether this extends to videos is, to our knowledge, yet to be shown.

Contextual features. Videos that stand out from the rest of the corpus (the set of images or videos from which the memorability experimental sequence is sampled) and therefore that differ from the observer's expectations are usually remembered better [35]. Additionally, when comparing the memorability of labelled scenes or objects, an item that belongs to an unusual category or whose attributes have a low frequency in the corpus will have a higher memorability [10, 28, 33, 35]. Lukavský and Děchtěrenko [35] also showed that computing these similarity measures on deep semantic features tend to be more predictive of memorability than using low-level descriptors, although these two approaches are actually quite complementary. Goetschalckx *et al*. [24] compared several distinctiveness metrics, including CNN-likelihood [10], sparseness [35] and asking participants to estimate the distinctiveness of images themselves, showing that all of these metrics do correlate well with memorability.

3.2. Predictive capacity of low-level descriptors

We first conduct a study of the predictive capacity of low-level descriptors of a video. These low-level descriptors are (i) *contrast*, defined as the standard deviation of the gray-valued representation of an image; (ii) *brightness*, defined as the mean of the hue channel of the HSV representation of an image; (iii) *blurriness*, as defined in [34]; (iv) *Histograms of Oriented Gradients* (HOG) [16], reduced to 10 components using PCA in order to mitigate overfitting. These image descriptors are computed on every frame of the video and averaged over the time axis to produce a unique scalar or vector feature. We also study: (v) the *mean optical flow* of the video, computed using OpenCV's TV-L1 implementation, averaged over each frame and over time; and (vi) the *size of the video* in bytes when resized to a shape of 256×256 and compressed using the H.264 standard.

We train a simple MLP with one 64-dimensional layer and a sigmoid activation on Memento10k [40] and VideoMem [13] to test the predictive capacity of each of the aforementioned features separately, once centered and reduced. As a baseline for our study, we use a constant model whose output is equal to the mean of the training and testing dataset (approximately 0.801 for Memento10k and 0.859 for VideoMem). Numerical results can be found in Tab. 1, and additional results can be seen in the supplemental.

The best low-level predictor is, unsurprisingly, HOG: it is the most advanced low level predictor in the set. On Memento10k, mean optical flow performs quite well, but not so much on VideoMem, possibly due to its observed lack of motion. On the other hand, the size of the compressed videos has a much stronger predictive capacity for VideoMem than for Memento10k. This could be connected to the more skewed distribution of movement in VideoMem, as the H.264 compression standard removes temporal redundancies. The combination of all factors unsurprisingly ends at the top position for both datasets, as feature fusion has been shown to improve empirical results in the memorability prediction literature.

Table 1. **Predictive capacity of raw descriptors.** We report the Spearman Rank Correlation ρ and the Mean Square Error (MSE) value between the ground truth memorability scores and the predictions of the MLP, that both give insights on the predictive capacity of the features.

	Memento10k [40]		VideoMem [13]	
Descriptor	$\rho \uparrow$	MSE $\downarrow$	$\rho \uparrow$	MSE $\downarrow$
All	0.383	0.0096	0.334	0.0058
HOG	**0.293**	**0.0103**	**0.311**	**0.0059**
Mean OF	$\underline{0.222}$	$\underline{0.0109}$	0.116	$\underline{0.0064}$
Contrast	0.112	0.0112	0.025	0.0065
Brightness	0.104	0.0113	0.098	0.0065
Size (bytes)	0.087	0.0113	$\underline{0.147}$	$\underline{0.0064}$
Blurriness	– 0.138	0.0114	0.136	0.0065
Baseline	–	0.0114	–	0.0145

4. Memorability modules and the pathway to memorability

We hypothesize that memorability prediction can improve if distinct modules could estimate the prevalence of each of the features defined in Section 3 individually. Additionally, the separation in low, mid, high and contextual features lends itself to the construction of a framework relating feature "tiers" to the final memorability of a visual input. We propose the concept of "pathway to memorability", a conceptual way of representing the hurdles that a visual input must overcome to achieve high memorability. We make the assumption that memorability is modulated by the presence and type of features exhibited by the visual stimuli, and that the modulation occurs through four feature encoders, f_{low}, f_{mid}, f_{high} and g, in the following way:

$$\begin{aligned} r &= f_{\text{low}}(x) \oplus f_{\text{mid}}(x) \oplus f_{\text{high}}(x) \\ m &= h(g(r, c) \oplus r)\,, \end{aligned}$$

where m is memorability, x is the visual stimuli, $\oplus$ is concatenation, g is the function that estimates contextual features, c is the context for x, f_{low}, f_{mid}, f_{high} are the functions that estimate the prevalence of low, mid and high-level features respectively, and h is the function that predicts memorability from concatenated features.

This representation allows us to think of memorability prediction as a combination of different factors, where each factor can be modeled individually. Based on this, we propose a new model architecture for video memorability prediction that instantiates specific networks for each function. We call this architecture Modular Memorability Model with

Similarity (M3-S): a model where each module (low, high and mid) contributes to the prediction, and similarity estimations connecting each module to a broader context are computed to take context into account.

5. M3-S: Modular Memorability Model with Similarity

We first propose the M3 model, a model that instantiates low, mid and high-level networks that compute representations for a given input, fuses them, and performs a regression to obtain a memorability score. The M3-S model corresponds to adding a contextual similarity module on top of this backbone (Fig. 1). The module takes one or more of the outputs of the previous modules, and yields a fourth feature vector, which is fused with the three others and fed to the memorability regression module.

This conceptual architecture can be instantiated with any set of modules that satisfy the concepts evoked in Sec. 4. In this section, we introduce a simple way of instantiating it using both perceptual and semantic descriptors — with an emphasis on the latter, as it has been shown that deep features tend to be more efficient than perceptual ones to predict visual memory performance [9, 18, 31, 43].

Modeling low-level video understanding. In order to target the main low-level factors of memorability, we explore the following set of image and video descriptors, defined and studied individually in Sec. 3.2: (i) *contrast*, (ii) *brightness*, (iii) *blurriness*, (iv) *HOG* reduced to 10 components. These image descriptors are computed on every frame of the videos and averaged over the time axis to produce a unique scalar or vector feature. We found that taking into account the temporal standard deviation did not improve the prediction capacity of our models. We also use: (v) *mean optical flow* of the video, and (vi) the *size* of the video in bytes when compressed using the H.264 standard. The output of the module is then the fusion of the outputs of the aforementioned descriptors. Although we showed in Sec. 3.2 that some features have low predictive scores when taken individually, performing a leave-one-out ablation study on their grouping gives evidence on the fact that we need them all to perform well on multiple datasets (see supplemental). We call this module *Raw perception*, as it estimates the contribution of the low-level perceptual factors to memorability.

Modeling mid-level video understanding. We choose a semantic segmentation network for the mid-level module. We use HRNetV2 [45], a recently proposed network that retains high resolution representations throughout the model. As a decoder for HRNetV2, we use a simple average pooling that reduces the output vector to a dimension $(B \times 720 \times 1 \times 1)$, where B is the batch size. We call this module *Scene parsing* as it detects objects and their distribution in the scene.

Modeling high-level video understanding. To model a high-level of understanding of the videos, we use the action recognition ip-CSN-152 network [47], whose ResNet backbone is stopped after the average pooling in order to yield a $(B \times 2048 \times 1 \times 1)$ feature vector. We call this module *Event understanding*.

Modeling contextual similarity. We perform the contextual similarity measure on both scene parsing and event understanding modules but not on the raw perception one (see Fig. 2), as it has been shown that semantic similarity is more predictive of memorability than perceptual similarity [10]. As shown on Fig. 6 and discussed in Sec. 6, these two learned representations are substantially different from each other and similarity in these feature spaces is very close to human semantic similarity, justifying the use of a similarity module that operates on them. As a means to evaluate how unusual a sample video is, we propose to cluster the feature space and utilize whether the video belongs to a cluster or not as an distinctiveness indicator.

We use the DBSCAN algorithm [19], as it relies on the density in the feature space, indicator of the commonness of a video sample in our dataset. For both scene and event modules separately, we reduce the dimensionality of the output features to 10 using PCA, then to 3 using t-SNE, before clustering the training output features. We then train a simple multi-layer perceptron (MLP) to perform label classification over the validation features, in order to obtain a label for each sample in the dataset.

In addition to using DBSCAN, we also explore a large number of similarity metrics, such as fractional distance [1], euclidean distance, cosine similarity and Kernel Density Estimation (KDE). We find that while some of these metrics yield distinctiveness values that can be reasonable predictors of memorability (distances from a feature to reference points in feature space allow to pinpoint the location of the feature), only DBSCAN significantly improves the performance of the M3 model. A summary of the results is provided in Tab. 2, and the full study can be found in the supplemental.

Predicting memorability from the features. To perform the memorability regression, we use a simple multi-layer perceptron (MLP) with two fully connected layers of size 512 and 64, using two Mish [37] and one sigmoid activation functions.

The detailed architecture of the instantiated M3-S model is depicted in Fig. 2. Additional details on the implementation can be found in the supplemental.

Figure 2. **Detailed architecture of the M3-S model.** The low-level descriptors used for the raw perception module are HOG, contrast, brightness, blurriness, mean optical flow, and video size in bytes. *For a model trained on Memento10k. The number of similarity features, *i.e*. the number of clusters given by the DBSCAN clustering, depends on the dataset used.

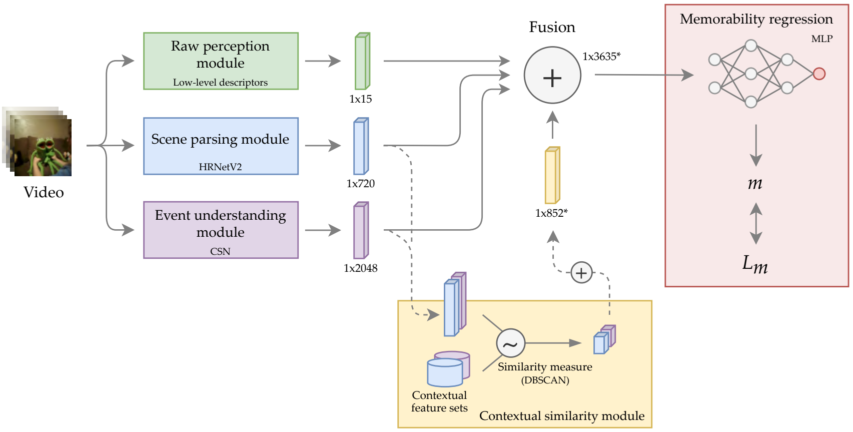
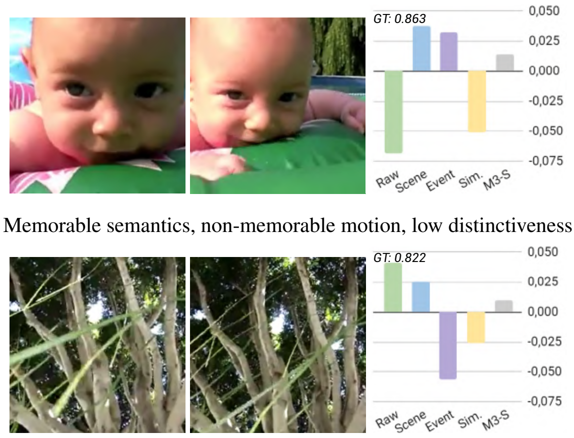


Table 2. **Predictive capacity of similarity measures** on the Memento10k dataset. We trained a simple MLP on 15 epochs to predict memorability based on the similarity features only. We leverage most similarity metrics with multiple different techniques and we report the results in this table. We define the *contribution* $\Delta\rho$ of a similarity measure s to the M3-S model as the difference between the Spearman score of the model using s and the M3 model, with no similarity module. For more details on the results and the techniques used, see the supplement.

Metric	$\rho\uparrow$	MSE $\downarrow$	$\Delta\rho\uparrow$
Fractional distance	0.491	0.00919	−0.0138
Euclidean distance	**0.514**	**0.00852**	−0.0021
Cosine similarity	0.321	0.01033	+0.0008
KDE	0.449	0.00928	−0.0025
DBSCAN clusters	0.340	0.01026	**+0.0053**



Memorable semantics, non-memorable motion, low distinctiveness.

Memorable motion, non-memorable event, low distinctiveness.

Figure 3. **Our M3-S model uses each level of the tiered representation** to produce precise memorability predictions. Here, we consider the predictions of the four modules alone (raw, scene, event, similarity) to separate their respective contributions, and compare them to the M3-S prediction. We report the deviation of these scores from ground truth. **Top:** The scene and event modules detect highly memorable semantics, like a human, but the raw and similarity modules detect a static and common clip and moderate these predictions. **Bottom:** Action semantics indicate a low-memorability video clip, while the raw perception module detects memorable dynamic patterns.

6. Experiments and results

We train our model and evaluate its performance on the two on the two main datasets for video memorability, Memento10k [40] and VideoMem [13]. Memento10k consists in 10,000 3-seconds "in-the-wild" clips taken from the Internet, encompassing a lot of variability in motion intensity and video quality, with additional captions that describe the content of the scenes; VideoMem consists in 10,000 7-second clips taken from professional video footage. Both datasets span a large semantic content diversity.

Previous work has showed that retraining is often unnecessary when it comes to visual memorability prediction, and that this approach was more likely to suffer from overfitting to the training set [41]. For this reason, we use pretrained weights on the scene and event modules, and only train the memorability regression MLP. We use a HRNetV2 pretrained on ImageNet [38], a CSN pretrained on IG-65M and fine-tuned on Kinetics-400 [20], and we use the Memento10k dataset [40] to train and evaluate our memorability predictor. We train our M3-S models for 20 epochs with a MSE loss and Adam optimizer, a weight decay of 10^{-5}, a batch size of 32, and a learning rate of 10^{-3} that decays by a factor 5 every 5 epochs. Because the most difficult samples are often in the tails of the ground truth memorability distribution, adding a penalization $p(m)$ in the loss function accounting for the distance of a ground truth memorability score m to the center of the distribution usually helps reaching a better performance. We approximate this by a simple penalization of the form $p(m) = \alpha|m - \bar{m}|^k$, where $\bar{m}$ is the average memorability of the dataset and $\alpha, k > 0$ are parameters that depend on the dataset. The loss function takes the form:

$$\mathcal{L}_1(m, \hat{m}) = \big[1 + p(m)\big] L_{\text{MSE}}(m, \hat{m}),$$

and we choose $\alpha = 2$ and $k = 4$ on Memento10k. We also train and test our model on the VideoMem dataset [13],

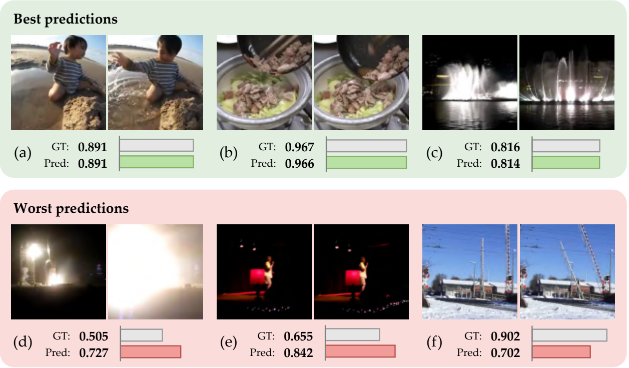

Best predictions
(a) GT: 0.891 Pred: 0.891
(b) GT: 0.967 Pred: 0.966
(c) GT: 0.816 Pred: 0.814
Worst predictions
(d) GT: 0.505 Pred: 0.727
(e) GT: 0.655 Pred: 0.842
(f) GT: 0.902 Pred: 0.702

Figure 4. **Best and worst predictions of M3-S.** Our model performs very well on scenes that involve a specific action **(a)** or a specific object **(b)** and on scenes whose semantic context is peculiar **(c)**. On the contrary, it can fail when given a video with strong color or scene variations **(d)**, a scene that contains a specific action but that is blurry or dark **(e)**, or a scene in which the uncertain outcome appeals to the viewer's anticipation **(f)**.

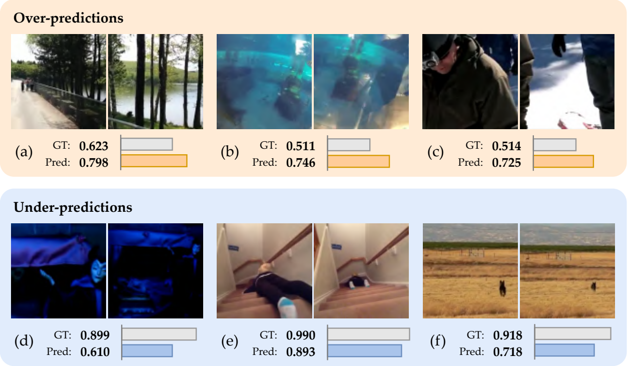

Over-predictions
(a) GT: 0.623 Pred: 0.798
(b) GT: 0.511 Pred: 0.746
(c) GT: 0.514 Pred: 0.725
Under-predictions
(d) GT: 0.899 Pred: 0.610
(e) GT: 0.990 Pred: 0.893
(f) GT: 0.918 Pred: 0.718

Figure 5. **Under and over-predictions of M3-S.** Our model overestimates the memorability of semantically bland scenes with humans **(a)**, of very dynamic scenes with no clear action **(b)**, and of scenes that contains memorable elements, such as humans or faces, but that are very shaky **(c)**, cluttered or blurry. Conversely, it underestimates the memorability of scenes that are emotionally salient — scary **(d)**, funny **(e)** — and of bland scenes containing a semantic content that is hard to grasp **(f)**.

for which we use a combination of the MSE loss and of the Spearman Rank Correlation loss L_{Spearman} proposed by Blondel *et al.* [8] that depends on the current training epoch $\text{ep} \in \{0, \dots, N_{\text{ep}} - 1\}$ through the multiplicative factor $\alpha_{\text{ep}} = \frac{\text{ep}}{N_{\text{ep}} - 1}$:

$$\mathcal{L}_2(m, \hat{m}) = (1 - \alpha_{\text{ep}}) L_{\text{MSE}}(m, \hat{m}) + \alpha_{\text{ep}} L_{\text{Spearman}}(m, \hat{m}).$$

To evaluate the performance of our model, we compare against prior work in video memorability prediction. We use the image memorability model MemNet [31] averaged over 7 frames of the video, the best performing video memorability models from Cohendet *et al.* [13], and the more recent work by Newman *et al.* [40]. We evaluate the models in term of Spearman Rank Correlation (RC), which is a popular metric [13, 29, 31] as rankings of memorability scores are more robust across experimental choices and external contexts. The results of our evaluations are in Tab. 3. Additionally, Figs. 4 and 5 show some prediction examples (best and worst cases as well as under and over predictions).

Table 3. **Comparison to state-of-the-art** on Memento10k and VideoMem, on which our approach, M3-S, surpasses the state-of-the-art in term of Spearman Rank Correlation ρ between the ground truth and predicted memorability scores. [*]As reported by [40]. [†]Without captions for fair comparison.

	Spearman RC $\rho \uparrow$	
Approach	Memento10k	VideoMem
MemNet baseline[*] [31]	0.485	0.425
Cohendet *et al.* (Semantic)[*] [13]	0.552	0.503
Cohendet *et al.* (ResNet3D)[*] [13]	0.574	0.508
SemanticMemNet[†] [40]	0.659	0.556
M3-S (ours)	**0.670**	**0.563**

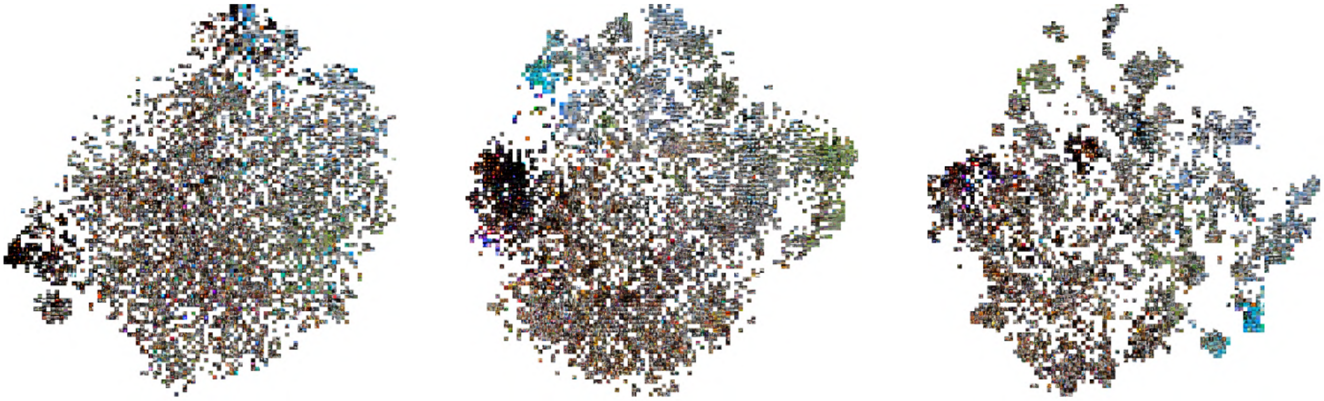
It is worth noting that the representations learned by our scene parsing and event understanding modules are substantially different from each other and that the similarity in the feature space is very close to the human perceptual semantic similarity. To demonstrate this, we perform t-SNE on the features of each module (raw, HRNet, CSN) and display a portion of the resulting image (Fig. 6). This justifies the use of a similarity module that operates on the feature space of the scene and event understanding modules, which is why we use the DBSCAN clustering algorithm. Details on DBSCAN clusters can be found in the supplemental.

Furthermore, we observe that the tiered representation allows each module to contribute to predicting a relevant memorability score using its specific level of video comprehension, as shown in Fig. 3.

7. Ablation Studies

We provide an in-depth ablation study of our M3-S model. The most significant study concerns the ablation of the different modules of the M3-S module. Its results can be found in Tab. 4 and the associated training curves are in the supplemental. We evaluate the models in term of Spearman Rank Correlation but also in term of Mean Square Error (MSE) so that we consider performance on both rankings and individual memorability values.

Out of the four modules of our M3-S architecture (raw, scene, event and similarity), the event module is the most crucial, and the importance of each module decreases with the level of understanding of the video it provides. This supports the idea that the memorability of a video lies first in its high-level semantics, and only then can low-level considerations — such as color, shape or motion — separate videos with the same semantics. For both datasets, the com-



(a) Low-level (HOG only)

(b) Mid-level (HRNet)

(c) High-level (CSN)

Figure 6. **t-SNE visualizations: each of our modules learns representations that are substantially different from each other. (a):** The raw perception module clusters clips based on shapes, colors, and motion. The t-SNE visualization of HOG features shows that video clips sharing similar shape properties are gathered by the HOG descriptor (top of the figure). **(b):** The HRNet module show a higher level of content understanding as it groups together clips with the same objects, even if they share different color and shape properties (*e.g.* vehicles, middle-right of the figure) **(c):** The CSN module goes even further and gathers clips sharing the same action semantics (*e.g.* swimming, bottom-right of the figure), even if they do not involve the same objects. See supplemental for parameters used to generate the figures.

Table 4. **Ablation study of our M3-S model.** We measure performance by computing the Spearman Rank Correlation ρ between the ground truth and predicted memorability scores, as well as the Mean Square Error (MSE), on both Memento10k and VideoMem. For each ablated version of M3-S, we also report the total number of features used by the MLP and its number of parameters. The number of similarity features, *i.e.* the number of clusters given by the DBSCAN clustering, depends on the dataset used.

	Memento10k [40]				VideoMem [13]			
Model	$\rho\uparrow$	**MSE** $\downarrow$	**Features**	**Parameters**	$\rho\uparrow$	**MSE** $\downarrow$	**Features**	**Parameters**
M3-S	**0.6699**	**0.00621**	3,635	1.89M	**0.5626**	**0.00469**	3,482	1.82M
M3-S without sim.	0.6303	0.00674	2,783	1.46M	0.5052	0.00476	2,630	1.38M
M3-S without raw	0.6249	0.00685	3,620	1.89M	0.4942	0.00483	3,467	1.81M
M3-S without scene	0.6139	0.00705	2,915	1.53M	0.4926	0.00489	2,762	1.45M
M3-S without event	0.5692	0.00779	1,587	0.85M	0.4433	0.00520	1,434	0.77M
Only event	**0.5988**	**0.00726**	2,048	1.08M	**0.4757**	**0.00493**	2,048	1.08M
Only scene	0.5295	0.00814	720	0.40M	0.4157	0.00525	720	0.40M
Only raw	0.3989	0.00950	15	0.04M	0.3321	0.00576	15	0.04M
Only sim.	0.3405	0.01027	852	0.47M	0.2267	0.00621	699	0.39M

bination of all modules gives the best performance, which shows that each level of understanding of the video plays a role in its overall memorability, and reinforces the relevance of a tiered approach to the video memorability prediction problem. Secondary ablation studies can be found in the supplemental.

8. Conclusion

Our contributions. We introduced a new paradigm for modeling memorability that instantiates separate modules focusing on key aspects of memory consolidation. We hypothesized that this modular formulation is competitive for memorability prediction, and our M3-S model confirms this hypothesis by surpassing the state of the art on two datasets. Our ablation studies show how each module contributes to memorability and shed light on the importance of a fragmented approach to this problem.

Limitations and future work. Video memorability prediction remains an open problem; Fig. 4 shows cases where our model fails to produce a good memorability score, often because of complex semantics, extreme pixel intensity or extreme motion. We believe that there is still room for understanding how to research each module, and that other instantiations could improve performance. One interesting possibility could be to overhaul the high-level module through emotion prediction; the bottleneck here appears to be the absence of competitive models and datasets for video emotion prediction decoupled from human faces.

References

[1] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001. 5

[2] Erdem Akagunduz, Adrian G Bors, and Karla K Evans. Defining image memorability using the visual memory schema. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2165–2178, 2019. 3

[3] Gal Almog and Yalda Mohsenzadeh. Memoir dataset: Quantifying image memorability in adolescents. 2021. 1, 2

[4] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323, 2013. 2

[5] Sathisha Basavaraju, Paritosh Mittal, and Arijit Sur. Image memorability: The role of depth and motion. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 699–703. IEEE, 2018. 3

[6] Sathisha Basavaraju and Arijit Sur. Multiple instance learning based deep cnn for image memorability prediction. *Multimedia Tools and Applications*, 78(24):35511–35535, 2019. 3

[7] Sathisha Basavaraju and Arijit Sur. Image memorability prediction using depth and motion cues. *IEEE Transactions on Computational Social Systems*, 7(3):600–609, 2020. 3

[8] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020. 7

[9] Zoya Bylinskii, Lore Goetschalckx, Anelise Newman, and Aude Oliva. Memorability: An image-computable measure of information utility. *arXiv preprint arXiv:2104.00805*, 2021. 1, 2, 3, 5

[10] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015. 1, 2, 4, 5

[11] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2018: Predicting media memorability task. *arXiv preprint arXiv:1807.01052*, 2018. 2

[12] Romain Cohendet, Claire-Hélène Demarty, and Ngoc QK Duong. Transfer learning for video memorability prediction. In *MediaEval*, 2018. 3

[13] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2531–2540, 2019. 1, 2, 4, 6, 7, 8

[14] Romain Cohendet, Karthik Yadati, Ngoc QK Duong, and Claire-Hélène Demarty. Annotating, understanding, and predicting long-term video memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 178–186, 2018. 3

[15] Mihai Gabriel Constantin, Bogdan Ionescu, Claire-Hélène Demarty, Ngoc QK Duong, Xavier Alameda-Pineda, and Mats Sjöberg. The predicting media memorability task at mediaeval 2019. In *MediaEval*, 2019. 2

[16] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 4

[17] Emmanuel Dellandréa, Liming Chen, Yoann Baveye, Mats Viktor Sjöberg, Christel Chamaret, et al. The mediaeval 2016 emotional impact of movies task. In *CEUR Workshop Proceedings*, 2016. 2

[18] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. What makes an object memorable? In *Proceedings of the ieee international conference on computer vision*, pages 1089–1097, 2015. 2, 3, 5

[19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. 5

[20] Facebook. Vmz: Model zoo for video modeling. https://github.com/facebookresearch/VMZ, 2018. 6

[21] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6363–6372, 2018. 3

[22] Alba Garcia Seco De Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Demarty Claire-Hélène, Faiyaz Doctor, Bogdan Ionescu, and Alan F Smeaton. Overview of mediaeval 2020 predicting media memorability task: What makes a video memorable? In *Working Notes Proceedings of the MediaEval 2020 Workshop*, volume 2882. CEUR Workshop Proceedings, 2020. 2

[23] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 3

[24] Lore Goetschalckx, Pieter Moors, Steven Vanmarcke, and Johan Wagemans. Get the picture? goodness of image organization contributes to image memorability. *Journal of Cognition*, 2(1), 2019. 3, 4

[25] Lore Goetschalckx and Johan Wagemans. Memcat: a new category-based image set quantified on memorability. *PeerJ*, 7:e8169, 2019. 2, 3

[26] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1633–1640, 2013. 3

[27] Junwei Han, Changyuan Chen, Ling Shao, Xintao Hu, Jungong Han, and Tianming Liu. Learning computational models of video memorability from fMRI brain imaging. *IEEE transactions on cybernetics*, 45(8):1692–1703, 2014. 2, 3

[28] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2013. 1, 3, 4

[29] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE, 2011. 1, 2, 3, 7

[30] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876, 2014. 1

[31] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398, 2015. 1, 2, 3, 5, 7

[32] Aditya Khosla, Jianxiong Xiao, Phillip Isola, Antonio Torralba, and Aude Oliva. Image memorability and visual inception. In *SIGGRAPH Asia 2012 technical briefs*, pages 1–4. 2012. 2, 3

[33] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. Relative spatial features for image memorability. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 761–764, 2013. 4

[34] Renting Liu, Zhaorong Li, and Jiaya Jia. Image partial blur detection and classification. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 4

[35] Jiří Lukavskỳ and Filip Děchtěrenko. Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception, & Psychophysics*, 79(7):2044–2054, 2017. 4

[36] Matei Mancas and Olivier Le Meur. Memorability of natural scenes: The role of attention. In *2013 IEEE International Conference on Image Processing*, pages 196–200. IEEE, 2013. 2

[37] Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019. 5

[38] MIT. Pytorch implementation for semantic segmentation/scene parsing on mit ade20k dataset. `https://github.com/CSAILVision/semantic-segmentation-pytorch`, 2019. 6

[39] Coen D Needell and Wilma A Bainbridge. Embracing new techniques in deep learning for estimating image memorability. *arXiv preprint arXiv:2105.10598*, 2021. 3

[40] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 223–240. Springer, 2020. 1, 2, 3, 4, 6, 7, 8

[41] Shay Perera, Ayellet Tal, and Lihi Zelnik-Manor. Is image memorability prediction solved? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3, 6

[42] Wen Qin and Chunshui Yu. Neural pathways conveying novisual information to the visual cortex. *Neural plasticity*, 2013, 2013. 2

[43] Sumit Shekhar, Dhruv Singal, Harvineet Singh, Manav Kedia, and Akhil Shetty. Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2730–2739, 2017. 2, 3, 5

[44] Hammad Squalli-Houssaini, Ngoc QK Duong, Marquant Gwenaëlle, and Claire-Hélène Demarty. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375. IEEE, 2018. 3

[45] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 5

[46] Lorin Sweeney, Mihai Gabriel Constantin, Claire-Hélène Demarty, Camilo Fosco, Alba G Seco de Herrera, Sebastian Halder, Graham Healy, Bogdan Ionescu, Ana Matran-Fernandez, Alan F Smeaton, et al. Overview of the mediaeval 2022 predicting video memorability task. *arXiv preprint arXiv:2212.06516*, 2022. 2

[47] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 5