EasyChair Preprint

№ 14970

Summarizing Video Content with a Single Image Using Large Language Models

Shutaro Tamada, Chunzhi Gu and Shigeru Kuriyama

September 21, 2024

Summarizing Video Content with a Single Image Using Large Language Models

1st Shutaro Tamada
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
tamada.shutaro.uv@tut.jp

2nd Chunzhi Gu
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
gu@cs.tut.ac.jp

3rd Shigeru Kuriyama
Department of Computer Science and Engineering
Toyohashi University of Technology
Toyohashi, Japan
sk@tut.jp

***Abstract*—Generating thumbnails for news videos plays an important role in efficiently understanding the contents. Prior techniques mostly handle this task by selecting one keyframe as a representative image. However, this approach cannot effectively handle a video whose key content is distributed across different frames. In this paper, we propose a summarization of a news video by composing its key contents into one image as a thumbnail. To achieve this, our method starts with text extraction from each scene in the video using OCR, speech recognition, and existing image captioning models. We then group these texts based on similarity and leverage large language models to score the group significance. Next, for each group, a keyframe is selected by jointly considering the importance and content quality. Eventually, we compose the objects in these keyframes into a single image as a thumbnail in a non-overlap manner and utilize diffusion-based generative models for further quality refinement. Experiments on real-world news videos demonstrate that our method can effectively extract key video contents and generate natural and informative video thumbnails.**

***Index Terms*—video thumbnail generation, video semantic analysis, large language models**

I. Introduction

Generating thumbnails for videos is an important task for people to understand the content quickly. For example, because there can be a large number of new events happening every day, watching the whole news video can consume a huge amount of time.

Prior efforts [1], [2] made for this task focus on selecting one keyframe that best summarizes the content for the whole video. Liu et al. [1] designed a multi-task deep visual semantic embedding model for video thumbnail selection. It considers two scores: a relevance score, which maps the similarity between queries and thumbnails into a common latent semantic space, and a representation score, which indicates how well the overall content of the video is represented. Despite the fact that it can generate the resulting thumbnail image, it cannot handle videos where important content is distributed in different frames. Because selection-based summarizing approaches tend to only consider the most significant frame, they tend to fail once the videos contain multiple key contents. To address this issue, Li et al. [2] proposed a method to interactively generate video thumbnails based on user queries. Differently, this method produces thumbnail candidates and generates one final result by pasting the clipped contents into one image. However, because the pasting is always performed in the same location, the results can appear unnatural.

In this study, we propose a method for composing multiple keyframes into a single image to provide a better solution to video thumbnail generation. In contrast to conventional approaches that select individual frames, our method generates the thumbnail to summarize the key content across the temporal domain. Our core idea is to assess the significance of the video content regarding text, audio, and semantics. To this end, based on the texts extracted from the videos, we employ large language models (LLM) to summarize the textual information on the image. We then perform grouping for the video frames according to the audio-visual similarity and select multiple keyframe candidates based on the estimated group importance. Finally, we segment the salient objects on each keyframe candidate to combine them into one result image with a diffusion-based generative model. Our model carefully evaluates the significance in terms of audio, visual, and textual information for keyframe determination. To improve the realism of the resulting image, we also assign the layout before composing the key objects. Experiments on 7 real-world news videos demonstrate that our method can effectively summarize the video contents into a single image with the awareness of multiple important objects.

II. Method

Given an input video $\mathcal{V}$, our goal is to generate a thumbnail image that well describes the content. Fig. 1 shows the schematic flow of our video thumbnail generation system. It involves two stages: (i) video semantics extraction (blue block) and (ii) thumbnail generation (red block). We detail each stage in the following section.

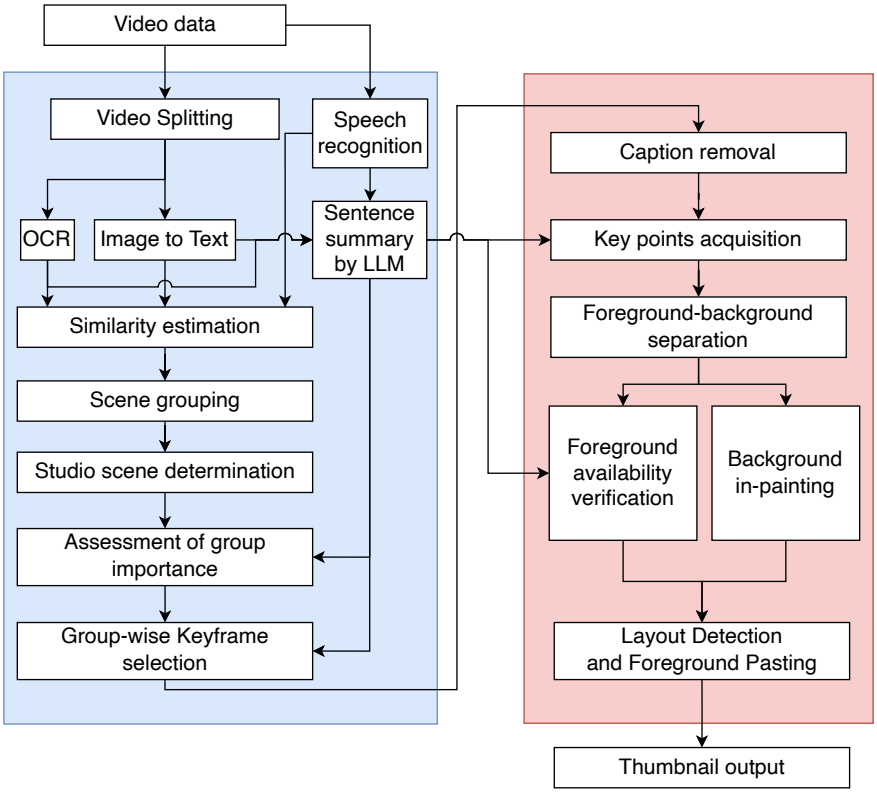


Fig. 1: **Flow of our video summarizing system.** The left (blue) stage depicts the scene-wise significance evaluation process, and the right (red) stage explains the key object positioning process to generate the thumbnail.

A. Scene-wise significance evaluation

Video scene splitting. Since the video content can be largely influenced by the scene it describes, we first need to perform scene segmentation to separate different semantic contexts. To achieve this, we use ContentDetector, which is a component of the PySceneDetect [3] that detects changes in scenes in the HSV color space by referring to the pixel changes in the HSV color space. For the input video $\mathcal{V} = [\mathbf{I}_1, \cdots, \mathbf{I}_t]$ with a total of t frames, We can thus obtain $N+1$ timestamps for camera switch times T_k^S, which allows us to divide the video into N segments $\mathcal{V}_i^S = \{\mathbf{I}_k | k = [T_{i-1}^S, T_i^S]\}$.

Text data extraction. A core factor in understanding the video content is the texts, including both audio and visual ones. To gain textual understanding, we next need to extract text from the video. Hence, we use Blip-2 [4], Speech Recognition (i.e., Whisper [5]), and Optical Character Recognition (OCR) to ensure a reliable reading of the information in each split video scene. In particular, Whisper can recognize multiple languages in the speech. Here, OCR is employed to read the captions/news footage in the scenes, and we use EasyOCR [6] used in CRAFT [7] and CRNN [8] in our implementation. Consequently, for the split video scenes, we obtain a three-fold text representation as $\mathcal{U} = [\{\mathbf{U}_1^{Aud}, \mathbf{U}_1^{Vis}, \mathbf{U}_1^{OCR}\}, \cdots, \{\mathbf{U}_N^{Aud}, \mathbf{U}_N^{Vis}, \mathbf{U}_N^{OCR}\}]$. Here, $\mathbf{U}_i^{\mathrm{Aud}}$ denotes the text obtained from speech recognition in the i-th scene, $\mathbf{U}_i^{\mathrm{Vis}}$ represents the text obtained from Blip-2 for the visual content, and $\mathbf{U}_i^{\mathrm{OCR}}$ indicates the text obtained from OCR of captions/news footage.

Similarity estimation and grouping. Using the extracted texts, we next need to determine the significance of each scene. Here, we use RapidFuzz [9] for efficient texture embedding acquisition. Since different scenes can share similar content, we group these scenes by referring to the similarity of textual embeddings. We prepare a weigh set $(\alpha^{Aud}, \alpha^{Vis}, \alpha^{OCR}) = (0.6, 0.2, 0.2)$ for similarity calculation, and then perform hierarchical clustering to obtain the grouped scenes $[\mathbf{S}_1, \cdots \mathbf{S}_G]$ with a total G groups, where each $\mathbf{S}_g$ includes one or many scenes.

News studio scene determination. We notice that news videos often include footage of the newscasters reading the lead text in the studio. The importance of such scenes tends to be higher than that of other scenes because the newscasters would read out the whole news content. In the context of a thumbnail image, which is designed to convey the essential content of the video, the studio scene is expected to be excluded from the candidate images to be used as thumbnails. To classify the in- and out-of-studio images, we train the Vision

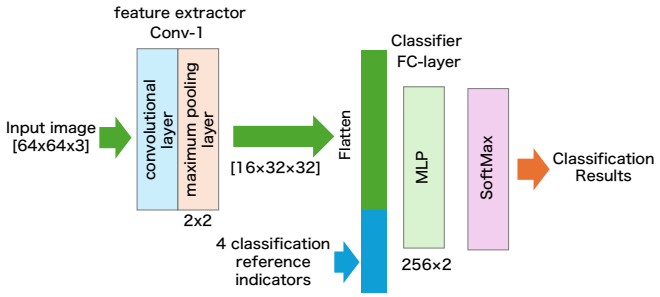


Fig. 2: **Architecture of our classification model to determine if an image can be used as a thumbnail.** Our model is structured with convolutional neural networks, which consider reference indicators in classification.

Transformer-based network [10] on 3500 studios and 8189 images from news videos to perform classification. This helps us exclude the in-studio scenes from thumbnail generation.

Group significance scoring. Because each scene group can have different importance levels in the whole video, we next need to score each group regarding the semantic similarity to the whole video. To achieve this, we leverage Phi-3-mini [11], which is a large language model (LLM) supporting fast and high-accuracy text processing. Specifically, we input each group text representation $\{\mathbf{U}^{Aud}, \mathbf{U}^{Vis}, \mathbf{U}^{OCR}\}$, and the entire texts $\mathcal{U}$ to the LLM to obtain the corresponding summarized texts. These texts are then fed to the Sentence Transformers [12] to infer the cosine similarity of each group regarding the entire video text. Eventually, we can derive the significance scores $[SC_1, \cdots SC_G]$ for each group, which we sort in descending order to facilitate further analysis. Note that since LLM only supports English, in tackling videos with other languages, we first translate them into English before using LLM.

Keyframe selection. Intuitively, each scene group can include multiple images with different semantic correlations to the whole video content. In addition, because news videos can include some unexpected factors, such as camera shakes or noise during interviews, some frames within the scene group can have low visual quality. We thus need to determine one keyframe that best represents the whole scene group. To achieve this, we train a convolutional neural network (CNN)-based classification network to determine whether an arbitrary frame can be used as a thumbnail candidate. The model overview is given in Fig. 2. During learning, we additionally inform the network with four reference indicators to better guide the classification: (i) semantic similarity between the whole image and the target image; (ii) score of camera shake status; (iii) area ratio of the subtitles; (iv) score of the human face-related factors.

For (I), we calculate the similarity of the summarized text of the video by LLM and the target frame by Blip-2, using

(a) Original frame (b) Caption-removed image

Fig. 3: **An example of Caption removal.**[1]

Sentence Transformer. For (ii), we apply the Laplacian filter to the image and compute the variance for all pixels. For (iii), we obtain the area ratio of the subtitles using YOLOv5 [13]. For (iv), we apply the Histogram of Oriented Gradients (HOG) with the Linear support vector machine (SVM) to detect facial regions from the image and regard the area of the largest bounding box normalized with the face number as the score.

To prepare the training data, we collected 2066 images by sampling from the videos and manually gave label annotations for all the images. The criteria for annotation are based on the perpetual recognition that judges the visual quality of a thumbnail. By using the trained classification network, we can select one keyframe for each group.

B. Thumbnail generation

Caption removal. As news videos can often include captions or subtitles, which are not desired in the thumbnail, we need to remove them before generating the result. To achieve this, we use the pre-trained YOLOv5 to detect the character areas and then directly remove them, including the bounding boxes or background areas. However, this would cause the images to have holes, which can greatly influence the generation. To address this, we use the Stable Diffusion 2.1 [14] to perform image inpainting to infill the holes. The positive text prompts in Stable Diffusion are obtained by using Image-to-Text results from Blip-2. Note that we employ LaMa [15], which is an image inpainting network using fast Fourier convolution, as pre-processing before using Stable Diffusion. This is because using Stable Diffusion only can sometimes damage the layout and destroy the semantic consistency. As such, we leverage Stable Diffusion as a post-processing to refine the results via LaMa. Fig. 3 shows an example of our caption removal.

Foreground-background separation. In many candidate thumbnail images, key objects of interest should be highlighted. It is, therefore, crucial to extract the foreground from the background and determine the keypoints of the object that we ultimately want to feature in the thumbnail.

To obtain the keypoints, we utilize both a Saliency Map and YOLOv5. The Saliency Map is derived using TranSalNet [16] and Smooth grad-cam++ [17]. For a given point of interest, we select the score obtained from the method with the higher Saliency Map score between these two techniques. By employing TranSalNet for local saliency prediction and Smooth grad-cam++ for global prediction, we can evaluate

[1]https://www.youtube.com/watch?v=1D6zxj8ONOA

(a) *Prosecutors search politician's office on suspicion of violating election laws.*[1]

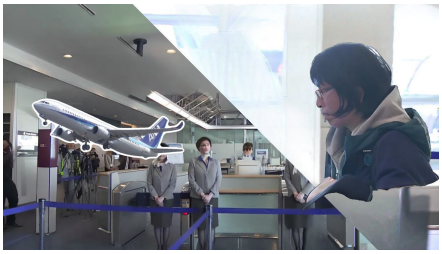
(b) *Flying Car Demonstration Experiment in Central Tokyo.*[2]

(c) *Bank of Japan Governor's comments on the weak yen.*[3]

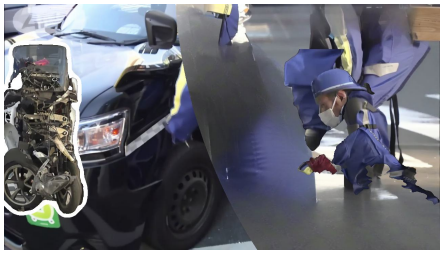
(d) *Akita Dogs Gather in Tokyo, Former Prime Minister Kan Comments.*[4]

(e) *Discussions and congressional references to constitutional amendments.*[5]

(f) *Politicians under search in alleged political funding scandal.*[6]



(g) *Passenger aircraft operations resumed to Noto Airport.*[7]



(h) *Cab and motorcycle collide at the intersection in Tokyo.*[8]

(i) *Tokyo Metro to Go Public; Tokyo and National Governments to Sell Shares.*[9]

Fig. 4: **Examples of the thumbnails generated by our method.** The captions for (a) $\sim$ (i) explain the textual description of the news videos.

saliency comprehensively. However, relying solely on the evaluation from the Saliency Map may lead to mistakenly detecting keypoints from the background. Therefore, we adopt keypoints only when they are detected as objects by the YOLOv5 pre-trained model, known for its general-purpose detection capabilities. Importantly, if more than one keypoints are detected, we compare the corresponding object categories produced by YOLO and the candidate object list summarized by LLM in the embedding space and adopt the one with the smallest distance as the keypoint.

Next, we use the U^2-Net [18] to separate the image into foreground and background based on the keypoints. Subsequently, we remove the parts corresponding to the foreground from the background and generate an image containing only the background. For foreground removal, similar to the caption removal case, we employ a combination of the LaMa and Stable Diffusion.

Layout arrangement. Our final step to generate the thumbnail is to paste the key objects in the image. This involves determining the layout of these objects so that they appear natural. In this step, we leverage significance to guide this process. For each group, we consider the frames with the top-2 significance score as "highly important", otherwise, "unimportant".

To improve the overall quality, we introduce a preprocessing step to exclude foregrounds where the video content summary by LLM and the text of the foreground object by Blip-2 differ significantly in the embedding space. Moreover,

[1]https://www.youtube.com/watch?v=Ojr-QqsPQC8
[2]ANNnewsCH /YouTube (Video already deleted)
[3]https://www.youtube.com/watch?v=1D6zxj8ONOA
[4]https://www.youtube.com/watch?v=M-VgGVMPhRQ
[5]https://www.youtube.com/watch?v=5yqdLYK9Dd8
[6]https://www.youtube.com/watch?v=YL3cRr6rToU
[7]https://www.youtube.com/watch?v=R9G-VHWHABU
[8]https://www.youtube.com/watch?v=0ZHRUVfPTmY
[9]https://www.youtube.com/watch?v=afN6dRJ5Xk4

(a) Officially provided thumbnail (b) Thumbnail generated by our method

Fig. 5: **Qualitative comparison with officially provided thumbnail images.** The news title is "*NATO secretary-general criticizes Trump's remarks.*"[1]

a foreground with greatly low saturation and brightness is also excluded.

For the "highly important" group, we first draw a line or circle in the background to divide the layout into two parts randomly. To ensure that each subject or critical area is effectively displayed, we perform several trials by different positions and orientations, and the final layout position is given by the values from the Saliency Map via TranSalNet [16] and Smooth grad-cam++ [17].

For the "unimportant" group, if the image can be clearly separated into foreground and background, we use the foreground. Otherwise, this group will be discarded.

After classifying the images into the "highly important" group, we proceed to arrange the images from the "unimportant" group. To do this, we systematically search for potential positions for each low-importance group image in a brute-force manner, starting from the least prioritized areas based on the saliency map of the high-importance group images. These positions should ideally not overlap with the foreground of the high-importance images. If we cannot find a suitable position for a low-importance image where it does not overlap with the foreground of any high-importance image, that particular low-importance image will be removed. This process ensures that only images that can be placed without obstructing the main subjects in the high-importance images are included in the final arrangement.

III. Experiment

In this section, we present experiments on real-world news videos to confirm the effectiveness of our method. Specifically, we select Japanese TV news from YouTube clips as the input video. Eventually, our experiments are performed on 10 videos.

We divide our evaluation protocol into two categories: (a) For news videos without actual thumbnails, we directly perform the qualitative inspection; (b) For news with provided official thumbnails, we evaluate the visual similarity to confirm the generation quality.

Evaluation of case (a). We present nine resulting images in Fig. 4, in which the original videos are selected to cover politics, social events, and industry. The news for Fig. 4(a) reports about the congressman's violation of election laws. We can see that the result seems reasonable as it presents the alleged politician and the prosecution officials investigating the case. Fig. 4(d) shows the output results for a news video about former Japan's Prime Minister Suga visiting and commenting on an event related to Akita dogs. The resulting images jointly include the Akita Dog and Prime Minister Suga, which is consistent with the news content. Interestingly, another former Prime Minister, Suga, also appears in the video. This is because our system jointly assesses textual and audio relationships to score the scenes, which gives the key person a high significance score.

Discussion. In Figs. 4(a)∼(g), the important objects are mostly placed without being covered by others. In Fig. 4(h), however, the car and the motorbike are overlapped, and the police officer is also not naturally placed. Also, the subway in Fig. 4(i) is not naturally placed.

The reason for such an unsatisfactory result can be attributed to the removal of the video subtitles/captions by YOLOv5 and the failure of realistic interpolation by LaMa and Stable Diffusion. We expect the solutions for these two issues to be (i) Fine-tuning the YOLO model with a larger amount of data and (ii) Providing Stable Diffusion with a more accurate text prompt. Also, the segmentation accuracy by U^2-Net could be potentially improved by providing multiple key points for guidance.

Evaluation of case (b). We next investigate the effectiveness of our method by referring to the actual news thumbnail for comparison. The result in Fig. 5 summarizes a video showing that the NATO secretary-general criticizes Trump's remarks. Although the layout assignment is different, our result successfully identifies the key persons: the NATO secretary-general and Trump, making the result resemble the actual thumbnail semantically. We can thus confirm that our proposed thumbnail systems achieve realistic results by comparing them with the true thumbnail, which demonstrates the effectiveness of our method.

IV. Conclusion

We have proposed a method for generating video thumbnails with a single image. To achieve this, we leverage large language models for textual extraction and score the semantic significance of the key objects to effectively understand their roles in the scene. The final thumbnail is generated by performing non-overlap pasting on a single image. Experiments on various real-world news videos generally demonstrate that our system is able to summarize the key content and produce a thumbnail that well describes the videos.

Our method is subject to several limitations, including excessive smoothing of image interpolation caused by stable diffusion and distortion of the extracted objects. We anticipate that addressing these limitations may be achieved by incorporating the camera angle as prior knowledge in scene cutting and by fine-tuning the stable diffusion model to more accurately represent real-world images. Additionally, we envision the potential extension of our method to summarize

[1]https://www.youtube.com/watch?v=tFPKvxU9DWs

diverse video formats. Our future work includes exploring these avenues for improvement.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers JP24K15247.

REFERENCES

[1] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3707–3715.

[2] J. Li, S. Lin, F. Zhou, and R. Wang, "Newsthumbnail: Automatic generation of news video thumbnail," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2022, pp. 1383–1388.

[3] Breakthrough, "Pyscenedetect," https://github.com/Breakthrough/PySceneDetect.

[4] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[6] JaidedAI, "Easyocr," https://github.com/JaidedAI/EasyOCR.

[7] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," 2019. [Online]. Available: https://arxiv.org/abs/1904.01941

[8] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," 2015. [Online]. Available: https://arxiv.org/abs/1507.05717

[9] RapidFuzz, "Rapidfuzz," https://github.com/rapidfuzz/RapidFuzz.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[12] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[13] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Z. Yifu, C. Wong, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain, "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," Nov. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7347926

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[15] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159.

[16] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "Transalnet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, 2022.

[17] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[18] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern recognition*, vol. 106, p. 107404, 2020.