TranBlend: An Experimental Fusion of Textual Input in Saliency Prediction Models

Benhao Huang, Shanghai Jiao Tong University

Abstract—Text-to-image models have revolutionized the application of AI-generated images (AGIs) across various fields such as design and entertainment. A particularly intriguing aspect of Computer Vision is Saliency Prediction—understanding where human focus naturally falls within an image. As of 2023, numerous models have shown remarkable performance on datasets like MIT300 and SALICON. However, human attention dynamically shifts with textual descriptions, suggesting a variance in saliency maps based on text context. This study explores whether AI can mimic this human trait. We focus on the SJTU-TIS dataset and introduce TranBlend Net, a model adept at incorporating text input into saliency map predictions, demonstrating the potential of AI in multi-modal understanding. Additionally, we delve into the loss function used for training the model, experimenting with a mix of various evaluation metrics to balance the dichotomy between overfitting and sufficient training.

Index Terms—Text-to-Image, AI-Generated Images, Saliency Prediction, MIT300, SALICON, TranBlend Net, SJTU-TIS, Multi-Modal AI, Loss Function, Model Training

I. Introduction

A. Background

saliency prediction, informed by human visual attention mechanisms, is a key task in computer vision. As Yan et al. [1] highlight, it focuses on predicting where human eyes are likely to be drawn in a scene. This task is vital for understanding and mimicking how attention is distributed across visual elements.

The goal of saliency prediction models is to replicate the human visual system's emphasis on distinctive features, such as color, contrast, or motion. These models are crucial in areas like user interface design, advertising, and autonomous driving systems, where pinpointing key elements significantly enhances performance and user interaction.

The development of these models has shifted from traditional methods that relied on manually crafted features to advanced deep learning techniques, which are fundamental to the methods used in this study.

B. Data Preparation

In the data preparation phase, SJTU-TIS [2] dataset encompasses five distinct categories of tasks: ”pure” images, which lack text annotations; ”general”(or called ”all”) images, accompanied by descriptions that encompass both salient and non-salient objects; ”salient” images, specifically annotated with descriptions of salient objects; ”non-salient” images, detailed with descriptions of non-salient objects; and ”whole” images, characterized by general scenario descriptions. A salient map serving as ground truth and fixation map used for calculating NSS score is provided for each image of each task type. For each data type, we adopt an 8:2 split, allocating 80% for training purposes and 20% for testing. This division is governed by a fixed random seed to facilitate reproducibility. Given that there are no prerequisites for data cleansing in our methodology, we have opted to bypass this step for the sake of clarity in our process.

II. Model Design

This section outlines the process of model design employed in our study.

A. saliency Prediction Model

In the realm of saliency prediction, numerous noteworthy models such as SAM [3] and salGAN [4] have established a strong foundation. These models predominantly leverage CNNs [5], LSTMs [6], and GANs [7] to effectively address saliency prediction tasks, demonstrating significant efficacy. However, a novel and inventive approach, TransalNet [8], has recently emerged. This model integrates a transformer architecture into the saliency prediction process, leading to remarkable improvements in performance, as evidenced by its results on the MIT300 [9] and SALICON [10] benchmarks. Therefore, attracted by its brilliant performance, we take TransalNet as a backbone model for saliency prediction part.

B. Text Features Extraction

Integrating textual instructions with visual data in model design presents a complex challenge. The objective is to enable the model to align text features with image features effectively. Addressing this challenge, BLIP [11] emerges as a well-suited solution. BLIP facilitates this alignment by implementing cross-attention mechanisms. Specifically, it combines text embeddings obtained from BertTokenizer [12] with image embeddings derived from ViT [13] within the BLIP architecture. This approach allows for a more nuanced and coordinated interplay between textual and visual information, essential for tasks requiring a deep understanding of both modalities.

Benhao Huang is an undergraduate student from Shanghai Jiao Tong University. E-mail:hbh001098hbh@sjtu.edu.cn

C. Feature Blending

In this experiment, the key challenge is how to effectively combine text features and image features to achieve the desired output from the model. A natural approach might be to integrate text features into the intermediate features of TransalNet, thus enabling the model to process text information. However, there's a size mismatch: the text features are of dimensions $(1, 1, 768)$, much smaller than the dimensions of TransalNet's middle layer features $(1, 768, 18, 24)$. This difference in size means that a direct blending of these features is not feasible.

To address this issue, we have tried several methods, including direct Up-sampling, Blending Net and Mask-Encoder.

1) Direct Up-sampling: To expand the dimensions of text features, a straightforward approach is to employ interpolation techniques such as 'Nearest', 'Bilinear', or 'Bicubic' for direct up-sampling. However, our experiments indicate that this method falls short for our specific task. The simplicity of direct interpolation, while appealing, fails to capture the complexity and nuanced interplay required between text and image features. Consequently, this naive up-sampling approach proves to be inadequate, necessitating the exploration of more sophisticated methods to effectively enhance the feature dimensions.

2) Blending Net: Given the limitations of direct up-sampling methods, we have identified Transposed Convolution as a superior alternative. This approach allows the model to learn up-sampling in a more flexible and effective manner. Upon obtaining two features of shape $(1, 768, 18, 24)$, we proceed beyond a mere element-wise dot product for blending these modal features. Instead, we concatenate them along dimension 1 and subsequently apply convolutional layers to this combined feature map. This method not only facilitates the integration of different feature modalities but also significantly enhances the model's learning capacity. By doing so, we aim to achieve a more sophisticated and effective feature blending, conducive to the model's overall performance.

3) Cross Attention Blending: In a similar way to how we derived text features from BLIP, we can employ cross-attention to enable the middle layer features of TransalNet to become aware of the text features from BLIP. This process involves using query, key, and value convolutional kernels, where the text features act as the query. They query against keys and values derived from the middle layer features of TransalNet. Essentially, this approach is akin to instructing the model to concentrate on the text features. Such a design is in line with our intuitive understanding and aims to foster a more meaningful interaction between text and image features, thereby enhancing the overall model performance.

III. Training Procedure

In this section, we will go through some training settings in our experiment. Some important hyperparameters are shown in table I. Throughout our training process, we consistently utilized the Adam Optimizer, maintaining static parameter settings. The batch size was set at 32 for all training sessions. In the initial phase for TranBlendNet, we employed a warm-up strategy, training the model across four types of tasks - "pure," "all," "salient," and "non-salient." This phase aimed to familiarize the model with the interplay between text and image features. Following this, the model underwent fine-tuning on each task type for an additional 15 epochs. To ensure the integrity of our evaluation, we strictly separated training and test data, reinforced by fixing the global random seed.For TransalNet, we adopted a similar approach, dedicating 25 epochs of training to each task, with identical optimizer and loss parameters as used in TranBlendNet.

A. How to Train

1) Model Warm Up: In our experimentation, we observed that TransalNet, having been thoroughly trained on the Salicon dataset, already exhibits excellent performance on tasks involving "pure", "salient", "whole" and "all" types of images. However, the most challenging and critical aspect of our experiment lies in enabling the model to recognize non-salient objects when prompted by specific texts. Consequently, we decided to primarily concentrate our training efforts on the "non-salient" dataset, aiming to enhance the model's responsiveness to textual information. For our blending net model, we first trained it on training dataset of four types task(pure, salient, all, non-salient) for 15 epochs, in order to have BLIP model warmed up on SJTU-TIS. For TransalNet, since it has no text input, it's obvious that there is no good for this warming up procedure, because the distribution of saliency map is actually biased by those text input, a pretraining on different types will only harm the performance of TransalNet on a certain task. As illustrated in Figure 4, TranSalNet's performance exhibits a decline during the later phase of training. This trend suggests that extending the training duration or conducting an extensive warm-up phase for TransalNet does not yield beneficial results. Instead, it may lead to decreased effectiveness, potentially due to overfitting.

2) Training on Non-salient Task: After necessary preparations, we started training the model for 25 epochs on non-salient type images to evaluate any improvements in test scores. A lack of improvement in this phase would indicate a need to further enhance the model's learning capabilities. Subsequent to this phase, we plan to fine-tune the model on other types of images, thereby ensuring a comprehensive adaptability to various image categories.

B. Loss Design

Mean Squared Error (MSE) loss is indeed an appropriate choice for tasks requiring pixel-level accuracy, as it effectively quantifies the discrepancies between the ground truth and the model's output. However, a significant drawback of using MSE loss is its tendency to cause model overfitting. This phenomenon is depicted in the figure 3, where rapid improvements in training performance

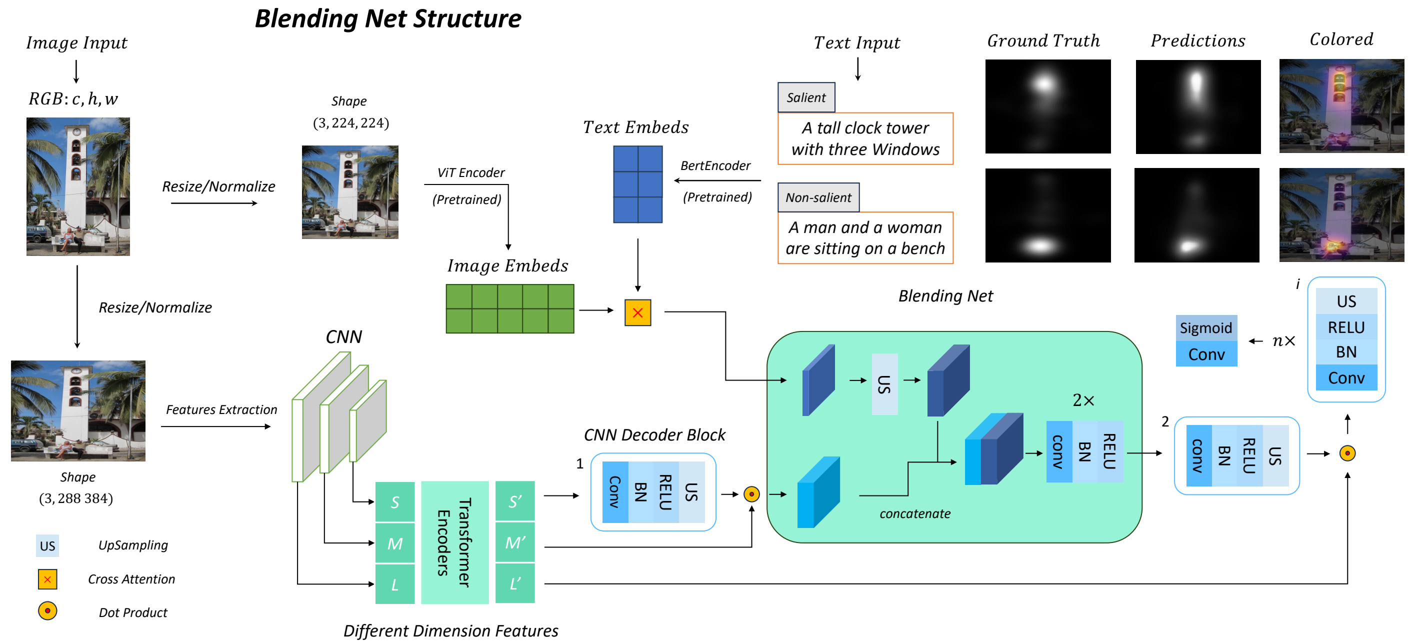

Fig. 1. The TranBlendNet Structure, depicted above, incorporates a multi-faceted approach to image and text processing. Initially, raw image inputs are resized to dimensions of $(3 \times 224 \times 224)$ for compatibility with the ViT Encoder from BLIP, and $(3 \times 288 \times 384)$ to align with the input requirements of TransalNet. For text inputs, we utilize the Bert Encoder to generate text embeddings. These embeddings then undergo a cross-attention mechanism with the image embeddings, facilitating a synergistic interplay between the textual and visual data. As outlined in TransalNet, a convolutional neural network (CNN) is employed to extract features at varying levels from the images. These features are subsequently amalgamated through a step-by-step dot product integration within the model's forward pass. A specialized blending block is crafted to merge the features derived from BLIP with the intermediate layer features from TransalNet. The resultant amalgamated features are then channeled through several CNN decoder blocks. The final output is produced after processing by a convolutional sigmoid layer, ensuring a refined and cohesive result. An illustrative example of this process is provided using an image from the test set, which is also included in the accompanying code files.

TABLE I

Key Hyperparameter Settings. In our notation, Fix Rate$_I$ represents the proportion of BLIP model layers that remained fixed during training. Similarly, Fix Rate$_{Te}$ corresponds to the fixed rate for the encoder module of TransalNet, and Fix Rate$_{Td}$ for its decoder module. Additionally, we implemented an 'AUC Threshold' to filter out pixels with low values, aiming to focus on the top 20% of pixels in terms of saliency.

Model	dataset	Training Epochs	Batch Size	Learning Rate	Adam Beta 1	Adam Beta 2	Adam Eps
TranBlendNet	SJTU-TIS	50/15	32	5e-06	0.9	0.999	1e-8

Model	dataset	Training Epochs	Batch Size	Learning Rate	Adam Beta 1	Adam Beta 2	Adam Eps
TransalNet	SJTU-TIS	25	32	5e-06	0.9	0.999	1e-8

w_{cc}	w_{nss}	w_{sim}	w_{kl}	w_{mse}	Fix Rate$_I$	Fix Rate$_{Te}$	Fix Rate$_{Td}$	Data Split Ratio	AUC Threshold
-2	-1	-1	10	4	0.9	0.0	0.0	8 : 2	0.8

are contrasted with declining test scores, when training TransalNet with merely MSE loss.

Here we provide some image examples from test data, which shows a clear overfitting patterns when merely using MSE loss for training.

To mitigate the effects of overfitting, we have adopted the loss function proposed in [8], which utilizes a combination of various metrics. These metrics include Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), Area Under the Curve (AUC), shuffled Area Under the Curve (sAUC), Kullback-Leibler (KL) divergence, and Similarity (SIM). Diverging from the original paper, our approach integrates a weighted Mean Squared Error (MSE) component into the combined loss function. This modification is designed to enhance the model's sensitivity to pixel-level differences between the actual saliency map and its predictions, while striving to minimize the risk of overfitting.

$$\mathcal{L} = \sum_{i \in S} w_i \cdot Score_i \tag{1}$$

where $S = \{cc, nss, kl, sim, mse\}$

Here w_is are hyper-parameters, which could be modified to achieve better performance.

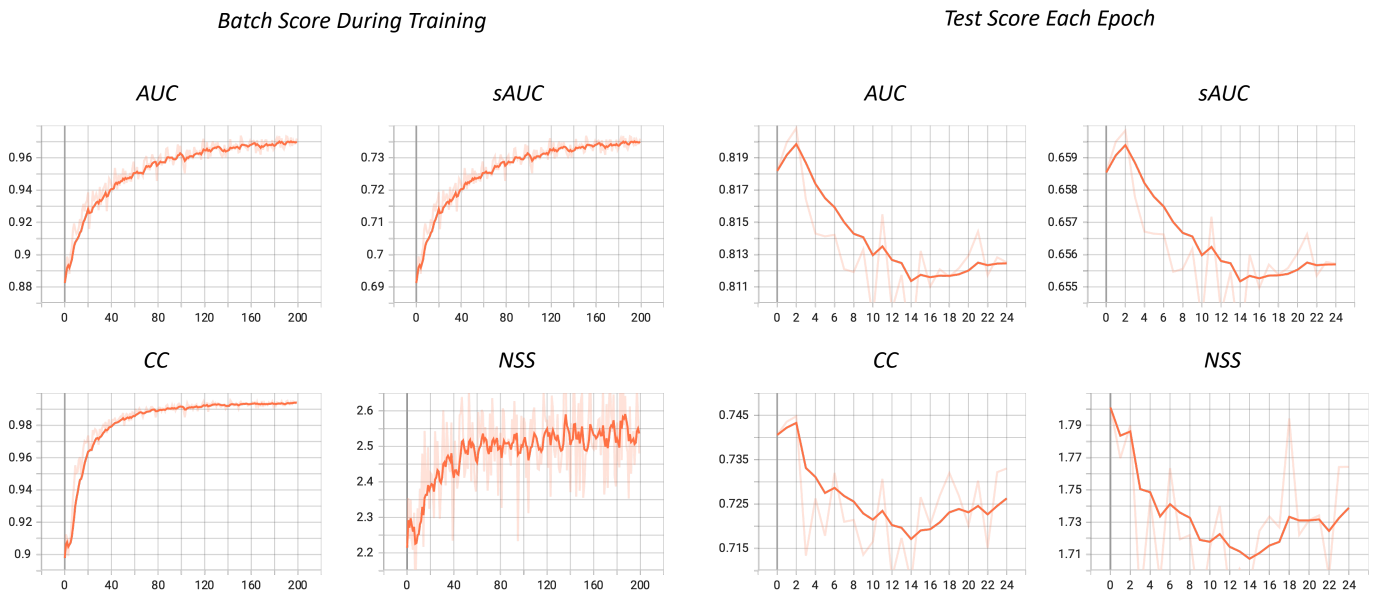

Fig. 2. Overfitting with MSE: The plot(smooth ratio = 0.8) shows results from training the TransalNet model for 25 epochs with 8 batches in each epoch. It clearly shows that even though the model gets better on the training data, it performs worse on the test data. This suggests that using MSE loss might cause the model to overfit. See figure 3 for specific image examples.

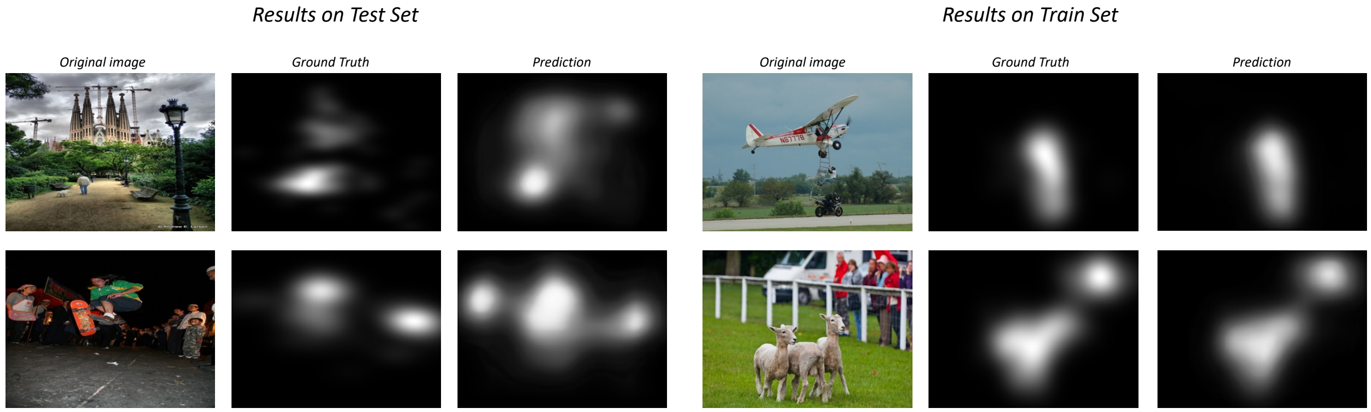

Fig. 3. As you can see from figures above, when merely trained with a MSE loss, the predictions made by models are almost no difference from ground truth on train set, but have a huge gap from ground truth when making predictions on test set. It seems that MSE loss can get quite excellent performance on training set, while perform poorly on test set. Therefore, we want to take best advantage of this metric, while eliminate its negative effects.

IV. Results and Analysis

A. Evaluate TransalNet on SJTU-TIS

In our initial evaluation, we assessed the performance of TransalNet on five tasks using the SJTU-TIS dataset. According to Lou et al. [8], there are two variants of TransalNet. We chose TransalNet-Dense due to its superior performance. Our experiments utilized a pre-trained checkpoint of TransalNet, which had been previously trained on the Salicon dataset[1]. Following this, we conducted a series of training and testing, spanning 25 epochs, with evaluations at each epoch. The results are presented below.

B. Focus on Non-salient Task

As illustrated in Figure 4, we observe that TransalNet demonstrates robust performance across all tasks, excelling particularly in the 'pure' task, which aligns with its original design. However, as anticipated, its performance on non-salient tasks is not as strong. Intriguingly, during training on the non-salient dataset, we noted a diverging trend: while TransalNet's performance consistently improved on the training set, it progressively declined on the test set. This is a typical indication of overfitting, where the model overly adapts to the training data at the expense of generalization to unseen data.

Then, we trained TranBlendNet mentioned in section II-C2 on non-salient dataset for contrast. As demonstrated in fig 5, it appears that TranBlendNet is apparently better

[1] https://github.com/ljovo/TransalNet

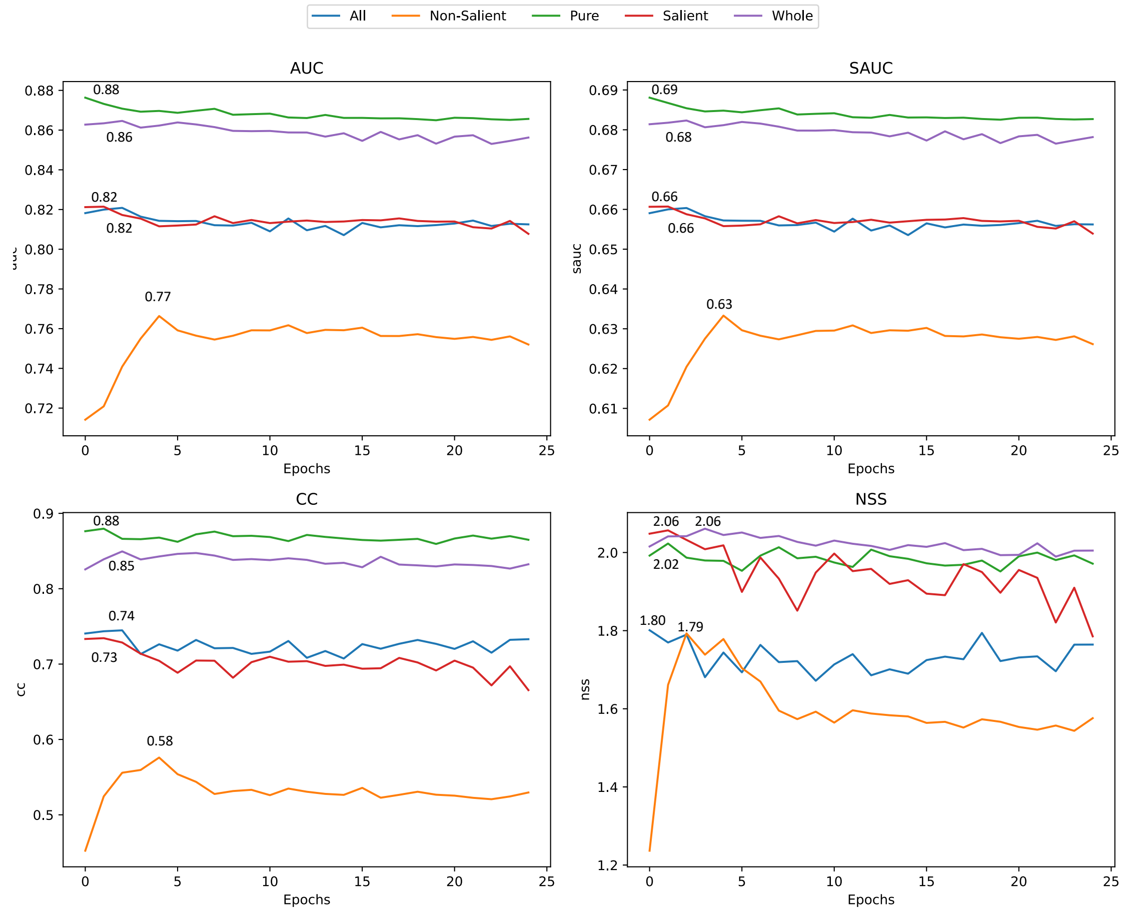

Fig. 4. The performance of TransalNet(Dense) on five tasks within the SJTU-TIS dataset, derived without text input. The results, as depicted in the accompanying figure, indicate that TransalNet excels on the pure dataset and performs admirably across other tasks. However, it exhibits less effectiveness on the non-salient task. This task, particularly challenging, demands a nuanced understanding of the interplay between text input and image input, a capability where the model shows limitations.

than TransalNet on non-salient task.

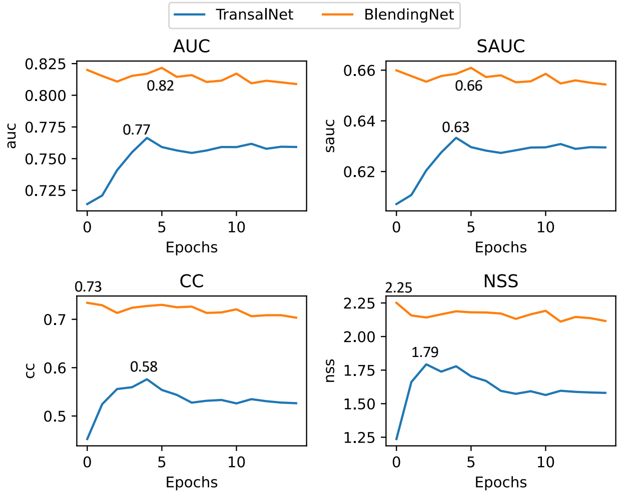

Fig. 5. Contrast between TransalNet and TranBlendNet on non-salient task. Both models carried out 15 epochs training on non-salient task, and scored for each epochs. The data of TransalNet used here is exactly the same with figure 4 of non-salient task. Therefore, an increased training epochs still cannot make TransalNet perform better on this task.

After going through an initial warm-up phase on four tasks, TranBlendNet seems to have picked up some useful patterns for identifying non-salient objects based on text input. During the training, its scores stay more or less the same, with just a small decrease that we think is due to overfitting. On the other hand, TransalNet starts by getting a basic understanding of the non-salient task dataset. But after that, it doesn't seem to learn much more. The scores actually go down as the training continues, which suggests it's struggling to improve after the initial learning.

We also tried cross-attention model, but it seems that cross attention on small scale is not that effective.

C. Ablation: Experiments on the fix rate of model

In the aforementioned experiments, while training TranBlend Net, we allowed the layers of TransalNet used in TranBlend Net to train freely from scratch, without any fixation. However, we observed a tendency in TransalNet to overfit the dataset. To gain a clearer understanding of this behavior, we conducted an ablation study specifically focusing on the non-salient dataset.

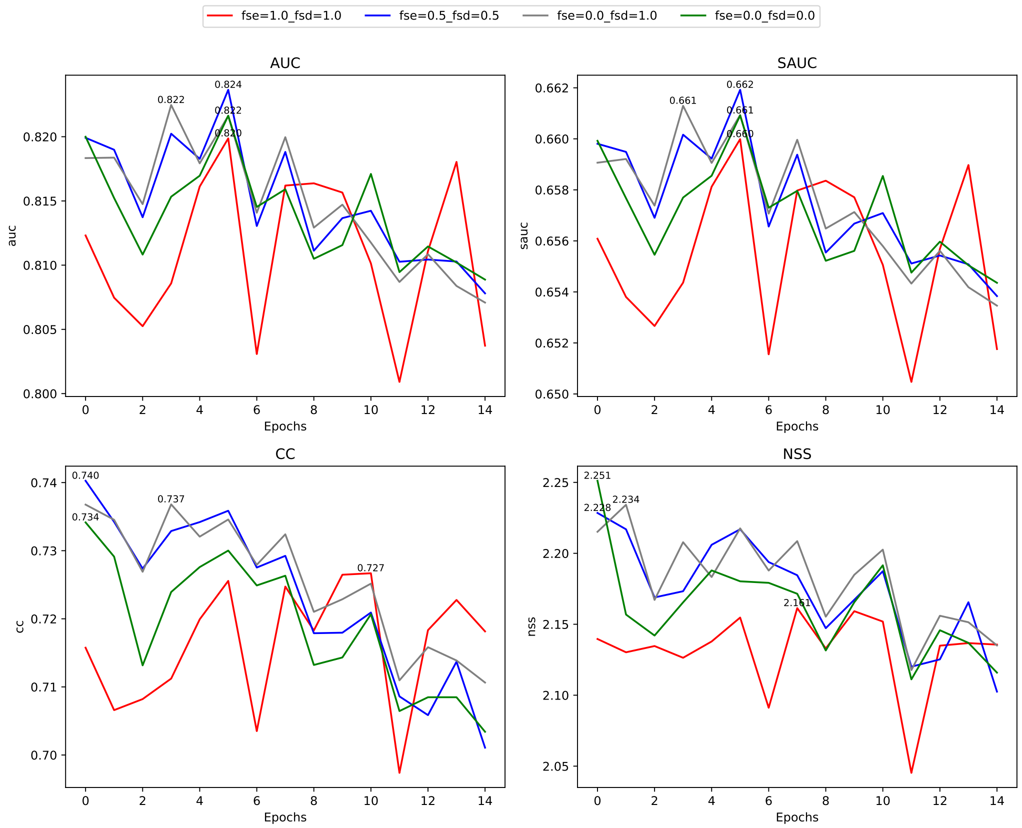

Fig. 6. In our ablation study, we examined the impact of varying fixation rates of TransalNet layers on overall model performance. While the performance fluctuations were within a relatively small absolute range, different fixation rates indeed had discernible effects on the model. As illustrated in the figure, completely fixing the TransalNet layers within TranBlend Net resulted in the poorest performance. This suggests that the TransalNet layers require some degree of adjustment in response to text inputs. Conversely, allowing too much flexibility can be counterproductive, potentially leading to overfitting. A moderate approach, involving fixing half of the TransalNet layers, yielded comparatively better model performance. However, given the marginal differences in outcomes and time constraints, we still present the results of the model with no fixed layers in Table II.

V. Final Results Represented in Table

In Table II, we present the highest scores achieved during the training epochs for both TranSalNet and TranBlendNet, with no layers fixed, across five different tasks using four key metrics: AUC, sAUC, CC, and NSS. This comparative analysis distinctly highlights the enhanced performance of TranBlend Net, particularly in the non-salient task, when contrasted with TranSalNet.

VI. Conclusion

In our study, we investigated the integration of text input with image features for text-guided saliency prediction. Our initial evaluation was conducted on the well-regarded TransalNet model using the SJTU-TIS dataset. We observed that while TransalNet excelled in ”pure” and ”salient” tasks, it underperformed in ”non-salient” tasks. We attribute this shortfall to its inability to effectively utilize the information encoded in text inputs.

To address this, we developed the TranBlend Model, which leverages the BLIP structure for handling text-image features. We utilized TransalNet as the backbone model, feeding its middle layer features, along with the text-image features from BLIP, into a blending block designed to encode the blended features. The results indicate that this method significantly enhances the Correlation Coefficient (CC) score in non-salient tasks and also shows improved performance in other tasks on the SJTU-TIS dataset.

Although our experimental setup had some limitations in rigor, the findings are promising. They suggest that by incorporating text information, we can guide models to focus on specific aspects of an image, marking a positive step toward the development of multi-modal models.

VII. Acknowledgment

The authors would like to thank the Teacher of and T.A of CS3324 for providing valuable advice which greatly help me complete this assignment.

References

[1] F. Yan, C. Chen, P. Xiao, S. Qi, Z. Wang, and R. Xiao, “Review of visual saliency prediction: Development process from neurobiological basis to deep models,” Applied Sciences, vol. 12, no. 1, p. 309, 2021.

TABLE II
Model Performance Across Tasks: It is evident from our analysis that TranBlend consistently outperforms in all tasks within the SJTU-TIS dataset. It's important to note that the scores presented here reflect each model's peak performance during the training process. This consideration is crucial due to the tendency of overfitting, as depicted in Figure 6, where we observe a slight decline in model performance when subjected to excessive training epochs. However, for fairness and consistency, all scores are derived from models trained for the same number of epochs. The best score of the column is in RED.

Task Type	Pure				General				Whole				Salient				Non-salient			
Model Name	AUC	sAUC	CC	NSS	AUC	sAUC	CC	NSS	AUC	sAUC	CC	NSS	AUC	sAUC	CC	NSS	AUC	sAUC	CC	NSS
TransalNet	0.8763	0.6881	0.8797	2.023	0.8208	0.6603	0.7448	1.801	0.8646	0.6823	0.8495	2.061	0.8214	0.6607	0.7344	2.057	0.7636	0.6319	0.5691	1.711
TranBlend	0.8757	0.6879	0.8889	2.087	0.8722	0.6862	0.8661	2.175	0.8742	0.6870	0.8692	2.089	0.8420	0.6711	0.8117	2.277	0.8200	0.6599	0.7342	2.251

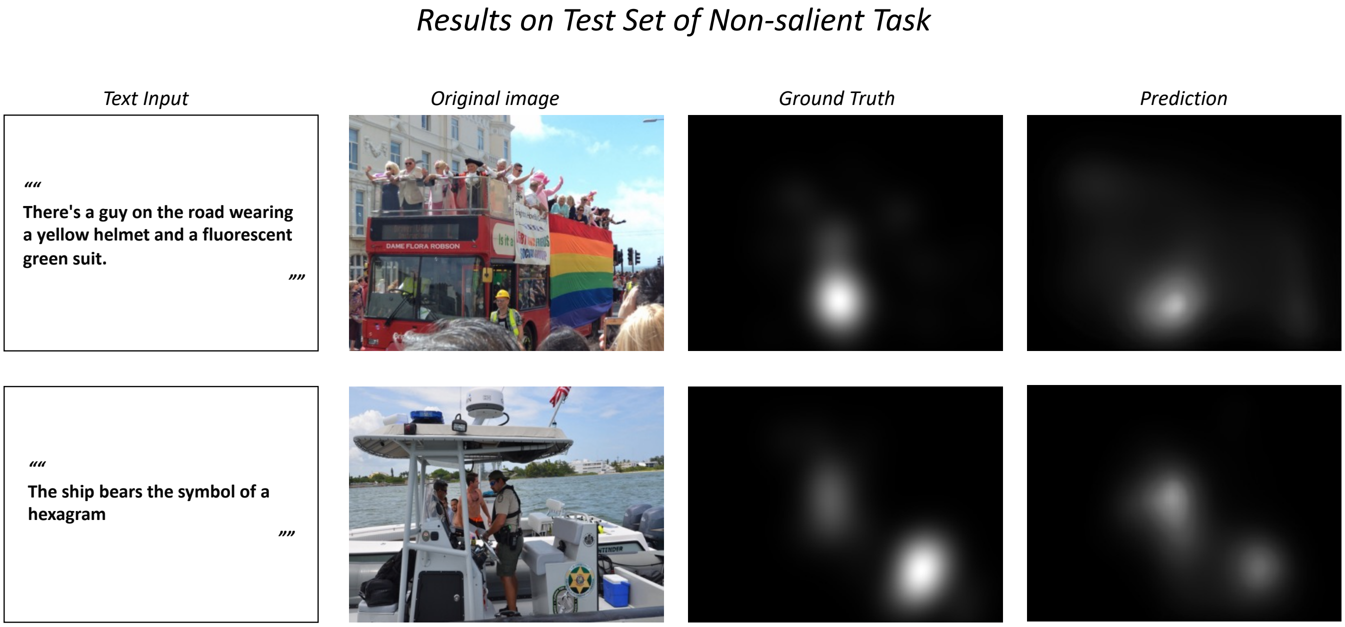
Results on Test Set of Non-salient Task
Text Input
Original image
Ground Truth
Prediction
"There's a guy on the road wearing a yellow helmet and a fluorescent green suit."
"The ship bears the symbol of a hexagram"

Fig. 7. Image Examples from TranBlend on Non-salient Task: Although TranBlend has not yet been able to perfectly replicate human attention patterns (as represented by the Ground Truth), it does demonstrate a certain ability to perceive and respond to text input.

[2] Y. Sun, X. Min, H. Duan, and G. Zhai, "The influence of text-guidance on visual attention," in 2023 IEEE International Symposium on Circuits and Systems (ISCAS), 2023, pp. 1–5.

[3] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Sam: Pushing the limits of saliency prediction models," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2018.

[4] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," arXiv preprint arXiv:1701.01081, 2017.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[8] 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.

[9] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark."

[10] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[11] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022. [Online]. Available: https://arxiv.org/abs/2201.12086

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[13] 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.

Benhao Huang is presently in his third year of undergraduate studies at Shanghai Jiao Tong University, majored in Computer Science. He holds a research internship position at the SJTU Interpretable ML Lab, where he works under the guidance of Prof. Quanshi Zhang. Recently, his research interests have been centered around Interpretable AI, Understanding Human Activity, Visual Reasoning, and Graph Learning. You can visit his website at huskydoge.github.io for more information. For any communication, he can be reached at hbh001098hbh@sjtu.edu.cn.