TDYSN: A YOLO-Based Top Down Saliency Prediction Network

Can Mızraklı* Tolga Çapın†
Department of Computer Engineering, TED University, Ankara, Turkey

*can.mizrakli@tedu.edu.tr †tolga.capin@tedu.edu.tr

Outline

Introduction

- Visual saliency identifies attention-grabbing regions.
- Bottom-up methods use low-level cues (contrast, edges).
- **Top-Down:** incorporate explicit task definitions (e.g., “count people”).
- We propose **TDYSN**: a lightweight model fusing YOLO features with Sentence-BERT embeddings via a transformer.

Key Contributions

1. TDYSN model: YOLO backbone + FPM + Sentence-BERT + Transformer fusion.
2. Trained on 1,968-image, four-task eye-tracking dataset: high performance with NSS AUC-Borji scores.
3. Thorough validation: quantitative metrics and qualitative examples.

Related Work

- Albayrak [1]: task-driven vs free-viewing saliency.
- VST [2]: transformer backbone for saliency detection.
- Murabito et al. [3]: classification-driven saliency.
- Simonyan et al. [4]: gradient-based saliency maps.

Dataset Preprocessing

- Dataset: 1,968 stimulus–FDM pairs across 4 tasks (Albayrak [1]).
- Split: 70% train, 15% val, 15% test (seed=42).
- Preprocess: resize stimuli to 384×384 and normalize to [0,1]; FDMs are 48×48, we down-sample predictions to match for the loss.
- Augment: paired horizontal flip (p=0.5), rotation ±10°.

TDYSN Architecture Overview

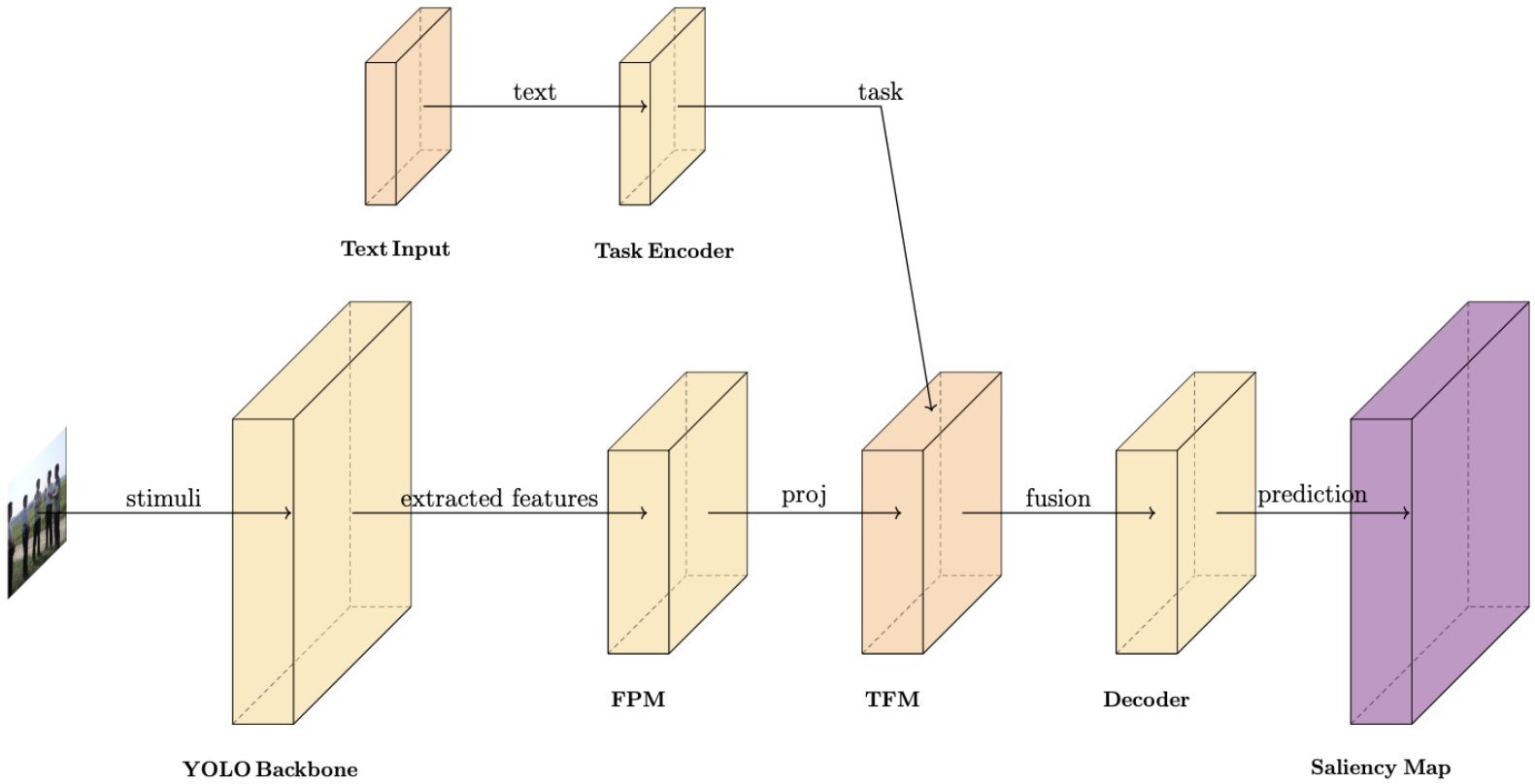

Figure 1: Model architecture overview.

Model Components

- **YOLO Backbone:** extract multiscale visual features.
- **Feature Projection Module:** 1×1 conv reduces 512→128 channels.
- **Task Encoder:** Sentence-BERT maps text task to 64-dim vector.
- **Transformer Fusion:** combines visual tokens + task token (d=128, heads=4).
- **Saliency Decoder:** Conv/Deconv upsamples to final saliency map (sigmoid).

Loss Function

$$\mathcal{L}(S, \hat{S}) = \alpha\, D_{KL}(S \| \hat{S}) + \beta\big(1 - \mathrm{CC}(S, \hat{S})\big)$$

- D_{KL} — Kullback–Leibler divergence: aligns the **probability distribution** of the prediction with ground truth.
- CC — Pearson correlation coefficient: measures **structural similarity** between predicted and true saliency maps.
- Hyper-parameters: $\alpha = \beta = 1$; predictions are down-sampled from 96×96 to 48×48 before loss computation.

KL and CC Formulas

$$D_{KL}(S\|\hat{S}) = \sum_i S_i \log \frac{S_i}{\hat{S}_i},$$

$$\mathrm{CC}(S,\hat{S}) = \frac{\sum_i (S_i - \bar{S})(\hat{S}_i - \overline{\hat{S}})}{\sqrt{\sum_i (S_i - \bar{S})^2}\sqrt{\sum_i (\hat{S}_i - \overline{\hat{S}})^2}}.$$

Training Configuration

- Optimization with the Adam optimizer (learning rate 1×10^{-4}) for stable end-to-end training.
- Loss combines Kullback–Leibler divergence (KL) and Pearson correlation (CC) to balance distributional and structural alignment.
- Trained over 40 epochs with regular validation checks to monitor for overfitting.

Performance Comparison

Model	NSS	AUC	CC	KLDiv	SIM
TDYSN (Ours)	3.53	0.9489	0.6433	0.9239	0.5112
EML-Net	2.05	0.8660	0.8860	0.5200	0.7800
Gold Standard	3.14	0.9341	0.9828	0.0602	0.8992

Table 1: Comparison with baselines.

Note: Higher CC/SIM in Gold Std. reflect smoother maps; task-conditioned focus decreases these metrics but improves NSS/AUC.

Training Loss Curve

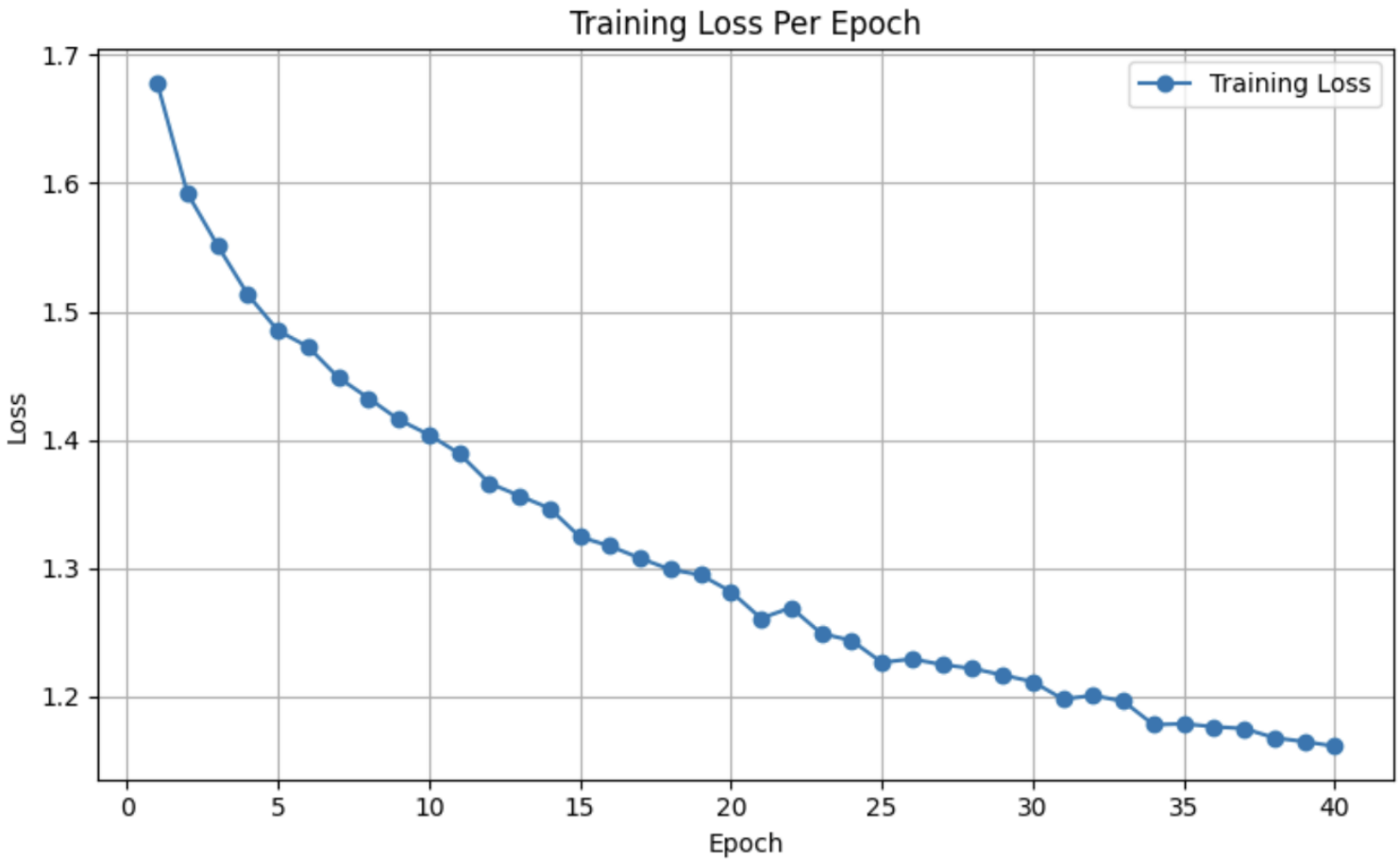

Figure 2: Training loss per epoch.

Validation Metric Trends

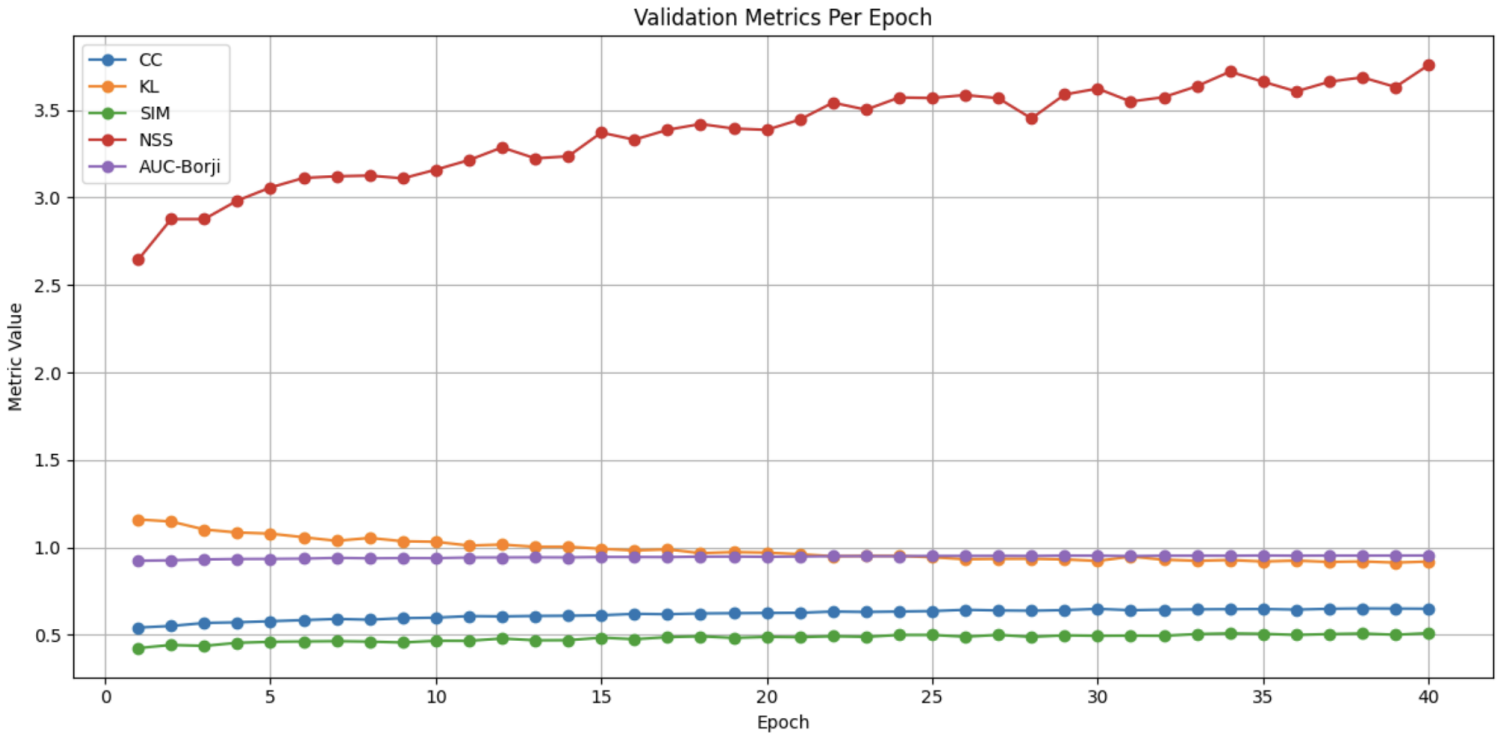

Figure 3: Validation metrics over 40 epochs.

Qualitative Visualization

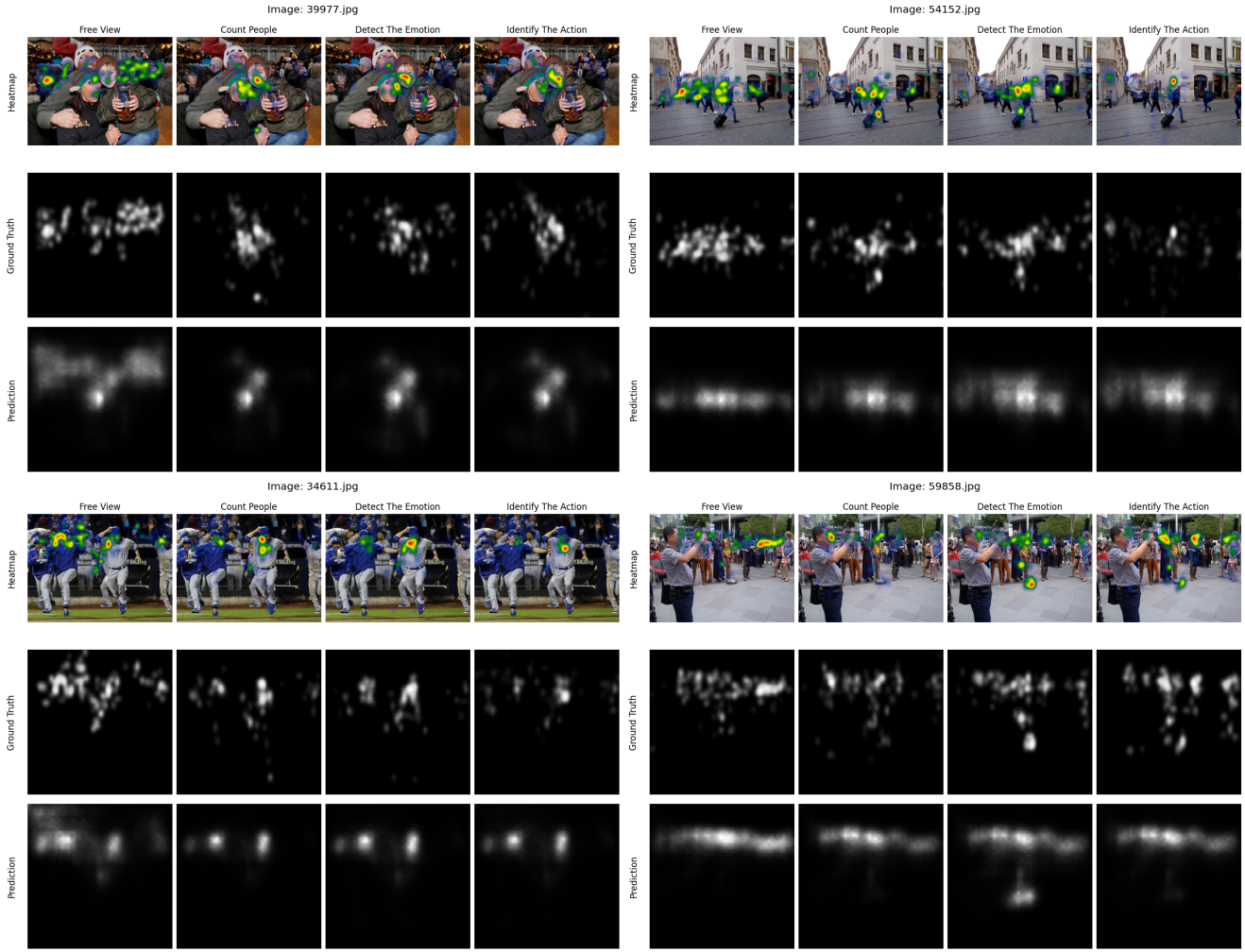

Figure 4: Saliency predictions for sample images.

Discussion

- **Interpretation:** NSS & AUC trends show stable task-driven focus.
- **Comparison:** Outperforms EML-Net & Gold Standard on NSS/AUC.
- **Limitations:** Dataset scale, no cross-benchmark, no ablation study conducted.
- **Future Work:** Data expansion, multi-layer fusion, planned ablation study.

Conclusion

- TDYSN: YOLO + BERT + transformer for top-down saliency.
- Performance comparable to successful models on task-driven dataset.
- Future: broaden data, cross-dataset validation, refine architecture.

References I


D. Albayrak, "A study of visual saliency for free-viewing and task-oriented condition," Master's thesis, TED University, 2020.


N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of ICCV*, 2021.


F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler, "Top-down saliency detection driven by visual classification," in *Proceedings of the Conference*, 2018.


K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv*, Dec. 2014.


Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2018.


M. Kümmerer, T. S. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.


——, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.


M. Kümmerer, T. Wallis, and M. Bethge, "Mit/tübingen saliency benchmark," 2024, https://saliency.tuebingen.ai/results.html.


J. Yu, Z. Li, M. Gong, W. Huang, and X. Shen, "Merw: A stochastic approach to saliency detection based on maximum entropy random walk," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2710–2717.


X. Jia and N. D. Bruce, "Eml-net: An expandable multi-layer network for saliency prediction," in *CVPR Workshops*, 2020.


J. Liang and Y. Zhang, "Top down saliency detection via kullback–leibler divergence for object recognition," in *Proceedings of the 4th International Symposium on Bioelectronics and Bioinformatics (ISBB)*. Beijing, China: IEEE, 2015, pp. 200–203.

References II

Z. Gniazdowski, "Geometric interpretation of a correlation," *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki*, vol. 9, no. 7, pp. 27–35, 2013.

A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, Jan 2013.

Thank You

Thank you for your attention!

Questions?