
Text-Guided Face Occlusion Segmentation using CLIP-based Cross-Modal Attention Fusion

Arpita Deshmukh
b23cm1007@iitj.ac.in

Sreenitya Thatikunta
b23cs1072@iitj.ac.in

Abstract

Segmentation of facial occlusions (e.g., glasses and face masks) is a key challenge in building robust face analysis systems or even face detection systems. Existing approaches are majorly vision-only which limits their ability to capture high-level semantic concepts that are naturally expressed in language. In this work, we introduce a cross-modal attention framework that integrates visual and textual representations to improve fine-grained occlusion segmentation.

Our approach leverages a pre-trained CLIP (Contrastive Language–Image Pre-training) model to bridge the gap between pixel-level visual features and semantic descriptions which mostly comes from text. Specifically, we reinterpret transformer attention maps as spatial priors and align them with text-derived representations, which helps the model to localize semantically meaningful regions corresponding to occlusions. We propose a Cross Modal Attention Fusion mechanism that facilitates structured interaction between visual attention and language embeddings, allowing the model to refine predictions using complementary cross-modal cues. This design enhances the model’s ability to segment the image properly in various challenging cases where visual signals alone are insufficient.

By incorporating language-guided attention into the segmentation process, this method help to incorporate semantic understanding into prediction tasks. This results in more robust and semantically consistent occlusion masks, demonstrating the potential of cross-modal reasoning for fine-grained visual segmentation.

1 Introduction

Face analysis is a fundamental component of applications such as biometric authentication and surveillance. A persistent challenge in these systems is handling facial occlusions, where regions of the face are partially covered by objects such as eyeglasses or masks. Such occlusions corrupt feature representations extracted by face encoders and can significantly degrade recognition performance. Accurately localising occluded regions as binary masks can therefore serve as an important pre-processing step for improving robustness.

Most existing approaches treat occlusion segmentation as a purely visual task. However, this overlooks the fact that occlusions are inherently semantic: categories such as “glasses” or “face mask” carry strong priors about shape, appearance, and spatial location. These priors are naturally encoded in language and can complement visual features, especially in settings with limited supervision. Recent vision–language models such as CLIP learn aligned representations across modalities which providing a natural foundation for incorporating such semantic understanding into dense prediction tasks.

In this work, we propose a framework that uses vision–language alignment to improve facial occlusion segmentation. Our key idea is to reinterpret internal attention representations from a pre-trained multimodal model as spatial priors, and to refine them through structured interaction with text-derived features. This enables the model to incorporate high-level semantic guidance while preserving

fine-grained spatial detail. By integrating language-aware attention into the segmentation process, this approach improves the model’s ability to localise the occluded regions.

2 Related Work

Transformer-based segmentation: Early work on image segmentation relied on fully convolutional networks (FCNs) (6), which enabled pixel-wise prediction directly from convolutional features. This idea was further improved by encoder–decoder architectures such as U-Net (5), which combine low-level and high-level features for better localisation. More recently, Vision Transformers (ViT) (2) have shown that attention-based models can perform as well as, or even better than, convolutional networks for visual tasks. Methods such as SETR (7) and Segmenter (12) extend Vision Transformers to segmentation by operating on sequences of image patch tokens and employing dedicated decoding mechanisms to produce dense pixel-level predictions. In our work, we also use ViT-based features, but instead of building a full decoder, we focus on using the CLS-token attention maps as a simple and effective spatial cue.

Attention maps as segmentation cues: Recent studies have shown that attention maps from transformer models often highlight semantically meaningful regions in images. For instance, DINO (3) demonstrated that self-supervised Vision Transformers can exhibit emergent object localization behavior, where attention maps roughly correspond to foreground objects despite the absence of labels. Similarly, Chefer et al. (8) proposed relevance-based attribution methods for transformers, which can be adapted to obtain spatial importance maps. Earlier approaches such as Grad-CAM (9) use gradient-based localization to identify important regions in convolutional networks. Inspired by these works, we use the CLS-to-patch attention from the final transformer layer as a coarse estimate of object regions, which is subsequently refined using a lightweight segmentation module.

Vision-language models for localisation. Vision–language models such as CLIP (1) learn a shared embedding space for images and text, enabling tasks like zero-shot classification and retrieval. Building on this idea, CLIPSeg (4) introduces a text-conditioned decoder for language-guided segmentation. Other works, including GroupViT (13) and OVSeg (19), extend this paradigm to open-vocabulary segmentation, allowing models to segment previously unseen categories using language supervision. Grounded-SAM (14) combines grounding models with segmentation systems to support flexible text-driven queries. In contrast to these approaches, our method is simpler and more focused: rather than relying on heavy decoding architectures, we fuse text features with visual attention maps using lightweight modules, resulting in a more efficient design that is well-suited for occlusion-aware segmentation.

Face occlusion handling. Handling occlusions in face analysis has been explored through a variety of approaches. Earlier methods include part-based models (11), which explicitly model individual facial components to account for partially visible regions, and mask-aware face recognition techniques (10), which aim to improve recognition robustness under occlusion. More recent work investigates generative approaches, such as inpainting, to reconstruct missing regions, as well as attention-based methods that reduce the influence of occluded features during representation learning. In contrast, our approach focuses on explicitly identifying occluded regions via segmentation, providing a complementary solution that can serve as a preprocessing step for downstream tasks such as recognition or reconstruction.

3 Occluded Dataset Construction

Training a reliable occlusion segmentation model requires a diverse set of face images along with accurate pixel-level annotations of occluded regions. Since no large-scale public dataset provides such annotations for common occlusions like glasses and face masks, we construct a synthetic dataset based on CelebA (16). The overall pipeline for dataset generation is illustrated in Figure 1.

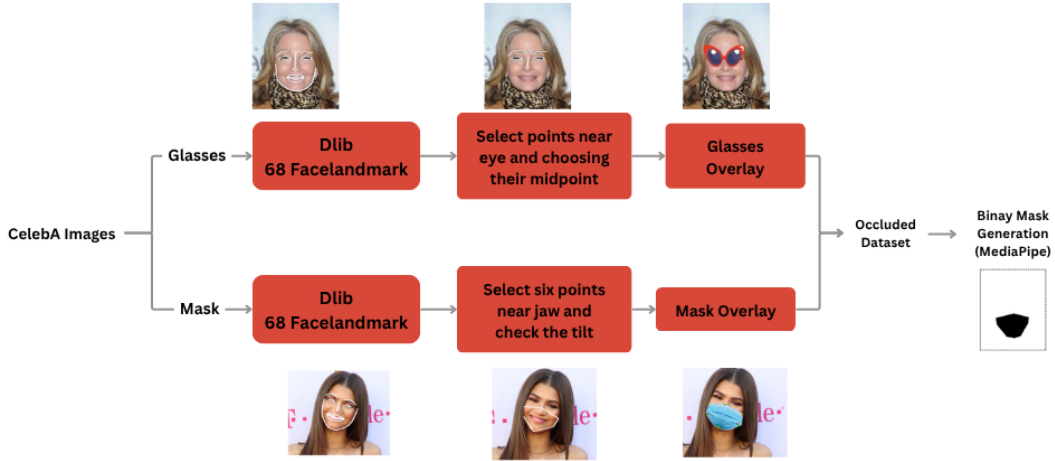


Figure 1: Pipeline for synthetic occlusion dataset construction. Starting from CelebA images, facial landmarks are detected, followed by occlusion augmentation (glasses or masks), and generation of corresponding binary segmentation masks.

3.1 Base Images

We use the CelebA dataset as the source of face images, which contains over 200,000 images of 10,177 unique identities, each having an average of 20 images. The images cover a wide range of poses and lighting conditions, making them a suitable base for generating realistic occlusions.

3.2 Synthetic Occlusion Augmentation

We introduce two types of occlusions: glasses and face masks.

- **Glasses occlusion:** Eyeglasses (including sunglasses) are added to images using OpenCV-based image processing. Facial landmark detection is used to identify the eye region, and the glasses are resized and positioned accordingly to ensure natural alignment with the face.
- **Face-mask occlusion:** Face masks are applied using the *MaskTheFace* framework (17), which provides realistic mask overlays based on facial landmarks. This ensures that the masks conform well to facial structure and appear consistent across different poses.

3.3 Ground-Truth Mask Generation

To generate ground-truth segmentation masks, we use MediaPipe (18) for facial landmark detection. The detected landmarks are used to define regions corresponding to occluded areas (such as the eye region for glasses and the lower face for masks). These regions are then converted into binary masks, where occluded pixels are marked as foreground.

3.4 Text Label Assignment

Each image is paired with a text label describing the type of occlusion. Instead of using a single fixed label, we define a small set of semantically similar prompts for each category. During training, one prompt is randomly selected for each sample to improve robustness to variations in phrasing.

- **Glasses:** “Glasses”, “Sunglasses”, “Eyewear”, “Covered eyes”, “Occluded eyes”
- **Face mask:** “Mask”, “Face mask”, “Covered face”, “Occluded region”, “Non-face region”
- **Clear face:** “Face”, “Human face”, “Visible face”, “Uncovered face”, “Clear face region”

4 Method

We propose a cross-modal segmentation framework that combines visual attention from CLIP with lightweight textual guidance. The central idea is to use attention maps as spatial priors and refine them using semantic information derived from language.

Given an input image and a corresponding text prompt (e.g., “glasses” or “face mask”), the image is processed by the CLIP visual encoder, while the text is processed by the CLIP text encoder. From these two branches, we extract attention-based features. The visual attention map provides a coarse localisation of relevant regions, while the text features provide semantic context. These two modalities are fused using cross-modal attention modules within a segmentation head to produce a final binary mask.

Figure 2 presents an overview of the proposed architecture.

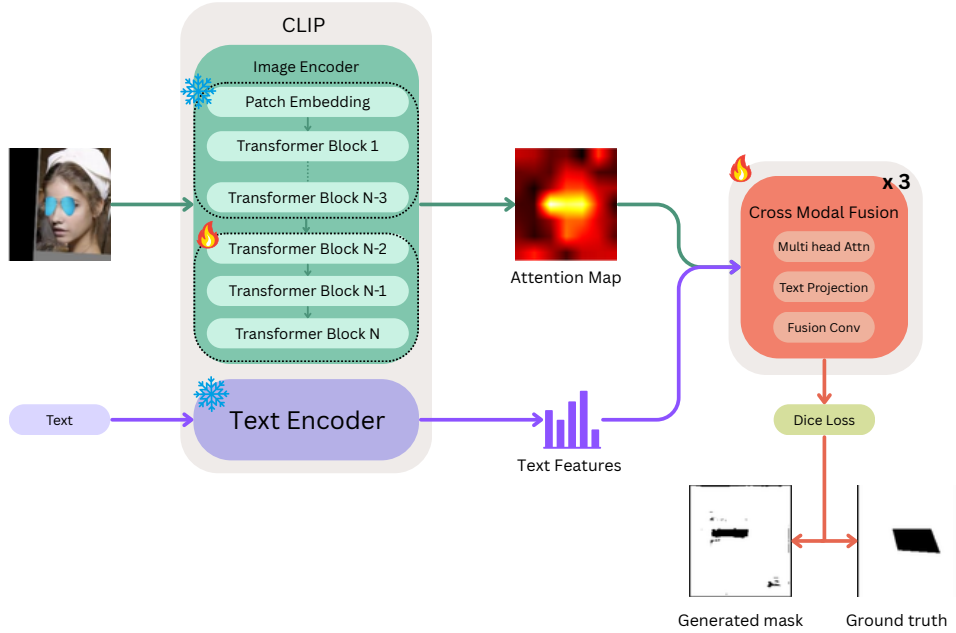


Figure 2: Overview of the proposed framework. Visual attention maps from CLIP are combined with text-derived features through cross-modal attention to produce a segmentation mask.

4.1 Backbone Design

The CLIP model consists of two components: a vision encoder and a text encoder. We adopt different training strategies for each:

- **Text encoder (frozen):** The text encoder is kept completely frozen during training. Since it is pretrained on large-scale image–text data, it already captures rich semantic relationships between language and visual concepts (e.g., “glasses”, “face mask”). Fine-tuning it on a relatively small, task-specific dataset may lead to catastrophic forgetting, where this general knowledge is degraded. Freezing the text encoder preserves its ability to generalise across different phrasings and unseen prompts.
- **Vision encoder (partially fine-tuned):** Only the final transformer blocks of the vision encoder are fine-tuned. Early layers primarily capture low-level features such as edges and textures, which are general and transferable across tasks. In contrast, deeper layers encode higher-level semantic representations, making them more suitable for adaptation to occlusion segmentation. This strategy allows the model to specialise while retaining the robustness of pretrained features, and also helps reduce overfitting and computational cost.

4.2 Role of Attention Maps

CLIP embeddings provide a global representation of the entire image, which is well-suited for classification tasks but insufficient for segmentation, where spatial localisation is required.

Attention maps, on the other hand, inherently preserve spatial information. The CLS token attends to different image patches, and the resulting attention weights indicate the relative importance of each region. We leverage these attention maps as an initial estimate of occluded regions.

4.3 Visual Attention Extraction

We extract attention from the final transformer layer of the visual encoder, specifically focusing on the interaction between the CLS token and image patches.

This produces an attention tensor:

$$\mathbf{a} \in \mathbb{R}^{B \times H \times N}$$

where H denotes the number of attention heads and N denotes the number of image patches. We average across attention heads and reshape the result into a spatial grid. This grid is then upsampled to the original image resolution:

$$\mathbf{A}_{vis} = \text{Upsample} \left(\frac{1}{H} \sum_h \mathbf{a}^{(h)} \right) \quad (1)$$

The resulting map provides a coarse spatial prior indicating regions of interest.

4.4 Text Feature Representation

From the text encoder, we extract a compact feature vector that summarises attention patterns across tokens. Instead of using the full embedding, we aggregate attention information from the final layers into a low-dimensional representation.

This feature vector serves as a semantic signal that guides the segmentation process.

4.5 Cross-Modal Attention Fusion

To combine visual and textual information, we employ a CrossModalAttentionFusion (CMAF) module.

Conceptually:

- Visual features indicate *where* relevant regions are located
- Text features indicate *what* the model should focus on

We use cross-attention to enable interaction between these modalities, allowing the model to refine visual attention using semantic cues:

$$\mathbf{V}' = \text{Fusion}(\mathbf{V}, \text{CrossAttention}(\mathbf{V}, \mathbf{f}_{text})) \quad (2)$$

4.6 Segmentation Head

The segmentation head refines the visual attention map through a series of convolutional layers interleaved with CMAF modules. The spatial resolution is kept fixed, while feature representations are progressively improved.

The final output is a single-channel mask with values in the range $[0, 1]$, representing the probability of each pixel belonging to an occluded region.

Table 1: Training hyperparameters.

Hyperparameter	Value
CLIP backbone	ViT-B/32
Image size	224×224
Batch size	24
Optimiser	AdamW
Learning rate	1×10^{-4}
Weight decay	0.01
Epochs	30
Train / val split	80% / 20%
Dice loss smooth ϵ	1.0
Text layers used	3
CMAF hidden dim	64
Seg. head hidden channels	128

4.7 Dice Loss

In occlusion segmentation, the foreground region (occlusion) is typically much smaller than the background. Standard losses such as binary cross-entropy treat all pixels equally, which can lead to suboptimal performance in such imbalanced settings.

Dice loss directly measures the overlap between predicted and ground-truth masks:

$$\mathcal{L}_{Dice} = 1 - \frac{2TP + \epsilon}{2TP + FP + FN + \epsilon} \quad (3)$$

Interpretation:

- The loss decreases as the overlap between prediction and ground truth increases
- It penalises both false positives and false negatives

Advantages:

- Focuses on region-level overlap rather than individual pixels
- Naturally handles class imbalance
- Encourages accurate and consistent segmentation boundaries

5 Experiments

5.1 Implementation Details

All experiments use the CLIP ViT-B/32 model loaded via OpenAI’s `clip` Python library. Images are pre-processed with CLIP’s default pipeline (resize to 224×224 , normalise). Masks are resized to 224×224 and binarised at threshold 0.5. Table 1 summarises all training hyperparameters. Training is performed on a single NVIDIA GPU (CUDA) and model checkpoints are saved every 2 epochs.

5.2 Ablation: Number of Frozen Vision Encoder Blocks

A central design question is how many of the twelve visual transformer blocks in ViT-B/32 to fine-tune. Freezing too many blocks limits adaptation capacity; fine-tuning too many risks overfitting and forgetting CLIP’s general visual representations. We ablate four configurations: freezing all but keeping the last 1, 2, 3, or 4 blocks trainable.

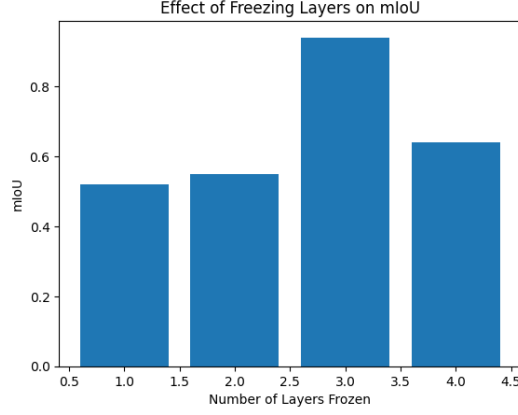


Figure 3: Validation mIoU as a function of the number of fine-tuned visual transformer blocks. Freezing more blocks preserves CLIP’s general representations but limits adaptation; fine-tuning too many blocks leads to overfitting.

5.3 Training Curves

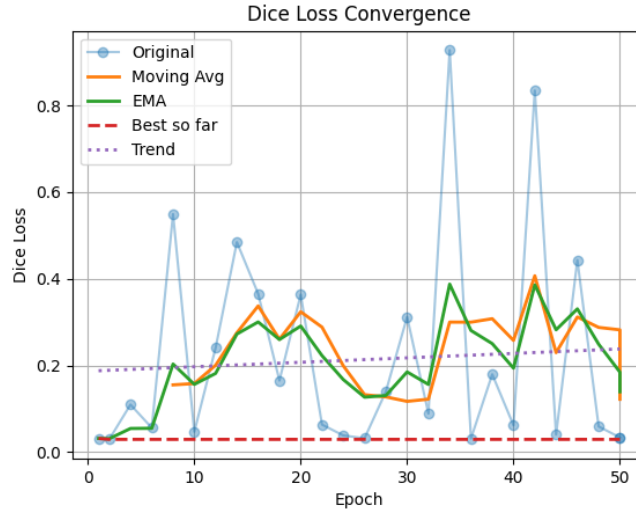


Figure 4: Validation Dice loss curves over 50 epochs for the selected configuration (last 3 visual blocks unfrozen).

The model shows a consistent downward trend in the validation Dice loss across training epochs (Figure 4), indicating that the cross-modal fusion head successfully learns to refine the CLIP attention maps into accurate occlusion segmentation masks. Table 2 reports Dice loss at selected epochs.

Table 3: Mean Intersection-over-Union (mIoU) scores for the trained CLIP-based segmentation model on held-out test sets.

Dataset	mIoU	Notes
CelebA (held-out)	0.9441	In-distribution
LFW	0.9405	Out-of-distribution

Table 2: Dice loss ($\downarrow$) at selected training epochs (last 3 blocks unfrozen).

Epoch	Train Dice Loss
6	0.0559
10	0.0468
18	0.1650
22	0.0633
24	0.0383
50	0.0336

5.4 Segmentation Results: CelebA and LFW

We evaluate the trained model on a held-out disjoint subset of CelebA and on the LFW dataset (15) to assess generalisation to an unseen distribution. Both sets are evaluated using the Mean Intersection over Union (mIoU) metric:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^N \frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \quad (4)$$

where P_i and G_i denote the predicted and ground-truth binary masks for sample i , respectively. Ground-truth binary masks for evaluation are generated using MediaPipe facial landmarks. Table 3 reports mIoU scores.

The model achieves an mIoU of 0.9441 on the held-out CelebA set and 0.9405 on LFW. The negligible drop between in-distribution and out-of-distribution performance demonstrates strong generalisation, attributable to the frozen CLIP text encoder’s robust semantic representations and the diverse prompt augmentation strategy employed during training.

5.5 Qualitative Analysis

Figure 5 shows example outputs of the predicted binary masks from the model and the attention maps that encode where the model focuses based on textual cues such as “masks” or “glasses”. Early in training, predictions are spatially diffuse and centred broadly around prominent facial features, consistent with raw CLIP attention maps attending to the overall face. As training progresses, the predictions sharpen and increasingly align with the ground-truth occlusion boundaries.

6 Discussion

Effect of freezing strategy. Keeping the text encoder frozen throughout training is critical: the CLIP text encoder has already learned precise semantic representations of occlusion categories, and fine-tuning it on a relatively small task-specific dataset risks degrading these representations. For the vision encoder, we find (Section 5.2) that unfreezing the last three blocks provides the best balance, consistent with the common finding that the later layers of vision transformers encode more task-specific features while earlier layers capture general low-level features.

Compact text representation. The text feature vector $\mathbf{f}_{\text{text}} \in \mathbb{R}^{B \times 3}$ is deliberately minimal: it encodes only the per-layer CLS attention statistics rather than the full 512-dimensional CLIP text embedding. This choice was motivated by parameter efficiency, but it represents a potential bottleneck. Richer text representations (e.g., the full CLIP CLS embedding) may further improve performance.

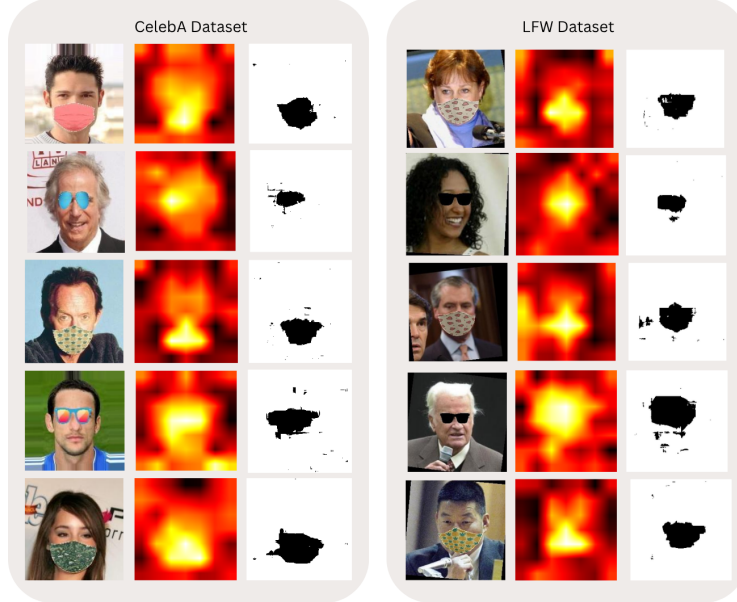


Figure 5: Qualitative segmentation results. **Column 1:** input image. **Column 2:** CLS-token attention map. **Column 3:** predicted binary mask.

Dice loss for class-imbalanced segmentation. Occlusion regions (glasses, face masks) typically cover a small fraction of the face image. In such class-imbalanced settings, binary cross-entropy loss can be dominated by the majority (non-occluded) class. Dice loss directly optimises the overlap between predicted and ground-truth masks, making it naturally robust to this imbalance. Our results confirm its suitability for this task.

7 Conclusion

We have presented a lightweight, language-guided face occlusion segmentation framework that repurposes the internal attention mechanisms of a pre-trained CLIP model. By combining CLS-token visual attention maps with compact multi-layer text attention statistics through a series of learned cross-modal fusion modules, the system achieves targeted segmentation of glasses and face-mask occlusions without requiring a heavy decoder or full backbone fine-tuning. Training with Dice loss is well-suited to the class imbalance inherent in occlusion segmentation. The model achieves an mIoU of 0.9441 on CelebA and 0.9405 on LFW, demonstrating both accuracy and generalisation. Systematic ablation over the number of fine-tuned visual transformer blocks confirms that unfreezing the last three blocks provides the optimal adaptation-regularisation trade-off.

Limitations and future work. Several avenues remain open. First, the text feature representation is very compact ($\mathbb{R}^3$); exploring richer encodings (e.g., the full 512-dimensional CLIP text embedding) may further improve segmentation quality. Second, the current approach uses prompt augmentation at training time but does not explore zero-shot generalisation to unseen occlusion categories; future work could evaluate on novel occlusion types. Third, incorporating data augmentation (random crops, photometric jitter) and more principled prompt engineering could improve generalisation. Finally, extending the framework to multi-class occlusion segmentation and evaluating on standard benchmarks such as the AR face database would provide broader validation.

Acknowledgments and Disclosure of Funding

The authors acknowledge the use of the CelebA dataset and the OpenAI CLIP library.

References

- [1] Radford, A., Kim, J.W., Hallacy, C., et al. (2021) Learning transferable visual models from natural language supervision. *ICML 2021*.
- [2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021*.
- [3] Caron, M., Touvron, H., Misra, I., et al. (2021) Emerging properties in self-supervised vision transformers. *ICCV 2021*.
- [4] Lüddecke, T. & Ecker, A. (2022) Image segmentation using text and image prompts. *CVPR 2022*.
- [5] Ronneberger, O., Fischer, P. & Brox, T. (2015) U-Net: Convolutional networks for biomedical image segmentation. *MICCAI 2015*.
- [6] Long, J., Shelhamer, E. & Darrell, T. (2015) Fully convolutional networks for semantic segmentation. *CVPR 2015*.
- [7] Zheng, S., Lu, J., Zhao, H., et al. (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *CVPR 2021*.
- [8] Chefer, H., Gur, S. & Wolf, L. (2021) Transformer interpretability beyond attention visualization. *CVPR 2021*.
- [9] Selvaraju, R.R., Cogswell, M., Das, A., et al. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *ICCV 2017*.
- [10] Wang, Z., Huang, Z. & Gong, S. (2020) Masked face recognition dataset and application. *arXiv:2003.09093*.
- [11] Song, L., Gong, D., Li, Z., et al. (2019) Occlusion robust face recognition based on mask learning with pairwise differential siamese network. *ICCV 2019*.
- [12] Strudel, R., Garcia, R., Laptev, I. & Schmid, C. (2021) Segmenter: Transformer for semantic segmentation. *ICCV 2021*.
- [13] Xu, J., De Mello, S., Liu, S., et al. (2022) GroupViT: Semantic segmentation emerges from text supervision. *CVPR 2022*.
- [14] Ren, T., Liu, S., Zeng, A., et al. (2024) Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv:2401.14159*.
- [15] Huang, G.B., Ramesh, M., Berg, T., & Learned-Miller, E. (2008) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *UMass Amherst Technical Report 07-49*.
- [16] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015) Deep learning face attributes in the wild. *ICCV 2015*.
- [17] Anwar, A. & Raychowdhury, A. (2020) Masked face recognition for secure authentication. *arXiv:2008.11104*.
- [18] Lugaresi, C., Tang, J., Nash, H., et al. (2019) MediaPipe: A framework for building perception pipelines. *arXiv:1906.08172*.
- [19] Liang, F., Wu, B., Dai, X., et al. (2023) Open-vocabulary semantic segmentation with mask-adapted CLIP. *CVPR 2023*.