Adaptive Depth Lightweight RGB-T Tracking with Holistic Token Routing

Supplementary Material

6. Similarity Metrics

6.1. Structural Similarity Index Measure

Structural Similarity Index Measure (SSIM) is a perceptually-motivated metric for evaluating the similarity between two images by comparing local patterns of luminance, contrast, and structure. Given two image patches $\mathbf{x} \in \mathbb{R}^{H\times W}$ and $\mathbf{y} \in \mathbb{R}^{H\times W}$, let $\mu_{\mathbf{x}}, \mu_{\mathbf{y}}$ denote their mean intensities, $\sigma_{\mathbf{x}}, \sigma_{\mathbf{y}}$ their standard deviations, and $\sigma_{\mathbf{xy}}$ their covariance. The SSIM is defined as:

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_{\mathbf{x}}\mu_{\mathbf{y}} + C_1)(2\sigma_{\mathbf{xy}} + C_2)}{(\mu_{\mathbf{x}}^2 + \mu_{\mathbf{y}}^2 + C_1)(\sigma_{\mathbf{x}}^2 + \sigma_{\mathbf{y}}^2 + C_2)},$$

where C_1, C_2 are small constants added for numerical stability. The output lies in the interval $[-1, 1]$, with 1 indicating perfect structural similarity. Owing to its strong correlation with human perception, SSIM is widely adopted in image quality assessment and is employed in our work to evaluate the structural fidelity of generated score maps.

6.2. Peak Alignment

Peak Alignment (PA) quantitatively evaluates the spatial consistency of the dominant responses between two score maps. For score maps $\mathbf{X}$ and $\mathbf{Y}$ of size $H \times W$, we extract their peak coordinates $\mathbf{p}_{\mathbf{X}} = \arg\max \mathbf{X}$ and $\mathbf{p}_{\mathbf{Y}} = \arg\max \mathbf{Y}$. The PA metric is defined as:

$$\mathcal{A}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\|\mathbf{p}_{\mathbf{X}} - \mathbf{p}_{\mathbf{Y}}\|_2}{\sqrt{H^2 + W^2}},$$

where the Euclidean distance between peaks is normalized by the diagonal length of the map. The output range is $[0, 1]$, with 1 indicating perfect alignment. This metric effectively captures localization consistency across network layers.

6.3. Cosine Similarity

Cosine similarity is a widely used measure of orientation alignment between two non-zero vectors in an inner product space. Given two feature vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, cosine similarity is defined as the cosine of the angle between them:

$$\text{CosineSimilarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|},$$

where $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$ denote the Euclidean norms of $\mathbf{a}$ and $\mathbf{b}$, respectively. The resulting value ranges from -1 to 1. Unlike distance-based metrics, cosine similarity is invariant to vector magnitude and focuses solely on directional agreement, making it particularly suitable for comparing feature representations in high-dimensional spaces.