Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion

Supplementary Material

1. Spatial Position Rules

Assuming the coordinates of the centroid pixel h_i are (x_i, y_i), we compare these coordinates based on positional descriptions to select the target instance that satisfies the given spatial relationship. We design 12 rules to address possible directional descriptions in the text. The first eight apply when the referenced object's position is easily describable, while the last four handle more complex cases, such as selecting an intermediate object among multiple instances:

1. Up: Select the instance with the highest position, i.e., the one with the largest y_i:

$$i^* = \arg\max_i \{y_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

2. Down: Select the instance with the lowest position, i.e., the one with the smallest y_i:

$$i^* = \arg\min_i \{y_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

3. Left: Select the instance that is furthest to the left, i.e., the one with the smallest x_i:

$$i^* = \arg\min_i \{x_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

4. Right: Select the instance that is furthest to the right, i.e., the one with the largest x_i:

$$i^* = \arg\max_i \{x_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

5. Top-left: Select the instance that is furthest up and to the left, i.e., the one with the smallest x_i and the largest y_i:

$$i^* = \arg\max_i \{y_i - x_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

6. Top-right: Select the instance that is furthest up and to the right, i.e., the one with the largest sum of x_i and y_i:

$$i^* = \arg\max_i \{x_i + y_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

7. Bottom-left: Select the instance that is furthest down and to the left, i.e., the one with the smallest sum of x_i and y_i:

$$i^* = \arg\min_i \{x_i + y_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

8. Bottom-right: Select the instance that is furthest down and to the right, i.e., the one with the largest x_i and the smallest y_i:

$$i^* = \arg\min_i \{y_i - x_i\}, \quad \forall i \in \{1, 2, \ldots, k\}.$$

9. N-th from the left: Select the Nth instance from the left by sorting x_i in ascending order:

$$i^* = \arg\min_{i \in S_N} \{x_i\}, \quad S_N = \sigma_\uparrow(x_i), \quad \forall i \in \{1, 2, \ldots, k\},$$

where $\sigma_\uparrow$ denotes the ascending function.

10. N-th from the right: Select the Nth instance from the right by sorting x_i in descending order:

$$i^* = \arg\max_{i \in S_N} \{x_i\}, \quad S_N = \sigma_\downarrow(x_i)x_i, \quad \forall i \in \{1, 2, \ldots, k\},$$

where $\sigma_\downarrow$ denotes the descending function.

11. N-th from the bottom: Select the Nth instance from the bottom by sorting y_i in ascending order:

$$i^* = \arg\min_{i \in S_N} \{y_i\}, \quad S_N = \sigma_\uparrow(y_i), \quad \forall i \in \{1, 2, \ldots, k\}.$$

12. N-th from the top: Select the Nth instance from the top by sorting y_i in descending order:

$$i^* = \arg\max_{i \in S_N} \{y_i\}, \quad S_N = \sigma_\downarrow(y_i), \quad \forall i \in \{1, 2, \ldots, k\}.$$

Finally, it is necessary to emphasize that our spatial relationships represent the relative rather than absolute positioning between multiple objects. We do not rigidly divide the image into fixed regions. For example, two people in the right area can be referred to as the "left" or "right person."

2. Architecture of Image Fusion Network

Holistically, our image fusion network receives multi-source images and text as input, and uses pseudo labels as supervisory signals. Specifically, this network consists of two stages: outputting the mask based on text localization instances and implementing two different fusion modes according to the mask.

First, we employ Restormer [10] blocks to extract features of the infared image I_{ir} and the visible image I_{vis}, as follows:

$$F_{ir} = E_{ir}(I_{ir}), \quad F_{vis} = E_{vis}(I_{vis}), \tag{1}$$

where E_{ir} and E_{vis} denote the infrared visual feature encoder and the visible light visual feature encoder, respectively. Next, the extracted visual features F_{ir} and F_{vis} are concatenated and passed through the visual feature fusion module $\mathcal{U}$ to obtain a unified visual feature V.

$$V = \mathcal{U}(F_{ir} \oplus F_{vis}), \tag{2}$$

where $\mathcal{U}$ represents the interaction operation via Restormer blocks [10]. Simultaneously, we utilize CLIP's text encoder with frozen weights to process the text t and obtain the text feature F_t, as follows:

$$F_t = CLIP_{text}(t), \tag{3}$$

where $CLIP_{text}$ denotes CLIP's text encoder.

Then, we construct an FPN (Feature Pyramid Network) using convolutional layers to enable multi-scale interaction between the text feature F_t and the unified visual feature V, ultimately yielding the positioning map P, i.e.,

$$P = \mathcal{S}(\mathcal{F}(V, F_t)), \tag{4}$$

where $\mathcal{F}$ represents the FPN, $\mathcal{S}$ represents the sigmoid operation. P is a binary image representing the location of the object referred to by the text. The white region indicates the ROI (Region of Interest), while the black region represents the non-ROI.

Subsequently, we utilize the positioning map P to perform Hadamard product with F_{ir} and F_{vis}, obtaining the feature representations of the ROI and non-ROI regions from the two source images, denoted by F_{ir_ROI}, F_{vis_ROI}, $F_{ir\_non-ROI}$, and $F_{vis\_non-ROI}$. Since different fusion modes are applied to the ROI and non-ROI regions, this step prepares for their independent fusion, preventing interference between the two regions.

$$F_{ir\_ROI} = P \odot F_{ir}, \tag{5}$$

$$F_{vis\_ROI} = P \odot F_{vis}, \tag{6}$$

$$F_{ir\_non-ROI} = (1 - P) \odot F_{ir}, \tag{7}$$

$$F_{vis\_non-ROI} = (1 - P) \odot F_{vis}, \tag{8}$$

where $\odot$ denotes the Hadamard product.

Next, we use Restormer blocks M_{ROI} to fuse the features of ROI regions, as follows:

$$F_{ROI} = M_{ROI}(F_{ir\_ROI}, F_{vis\_ROI}). \tag{9}$$

For non-ROI regions, we also use Restormer blocks $M_{non-ROI}$ but without shared weights, as follows:

$$F_{non-ROI} = M_{non-ROI}(F_{ir\_non-ROI}, F_{vis\_non-ROI}). \tag{10}$$

Finally, we input F_{ROI} and $F_{non-ROI}$ into the image reconstruction decoder D which consists of Restormer blocks to obtain the final fused image F:

$$F = D(F_{ROI}, F_{non-ROI}). \tag{11}$$

Clearly, we avoid directly computing the loss between the pseudo-labels and P; instead, we use them as a positional reference. Through supervision from two loss functions during training (Eq. X-Eq. X in the main text), the network is forced to learn: (1) accurate localization based on the text, i.e., generating P that closely matches the pseudo-labels, and (2) correctly applying the two fusion modes. Therefore, during inference (fusion), our model only requires the image fusion network, without needing the stage I model.

3. Why is our multimodal feature alignment module necessary?

Although there are many established multimodal feature alignment methods, they are not fully applicable to our task due to the following reasons:

Task-Specific Requirements: Existing methods, such as those based on cross-modal attention mechanisms [6], canonical correlation analysis [3], and contrastive learning [7], typically aim to align global features or focus on aligning image and text representations in a general sense. However, our task requires fine-grained, instance-level alignment where specific object instances need to be localized and highlighted based on user input. Most existing methods fall short of addressing this level of specificity.

Manifold Structure of Multimodal Data: Previous methods generally assume a linear or simple metric space for alignment, but multimodal data, particularly in the case of image and text, often exist in complex, nonlinear manifold spaces [4, 8]. These methods fail to fully exploit the manifold structure that underlies multimodal interactions. In contrast, our approach leverages manifold similarity as a guiding prior, which captures the intrinsic geometric relationships between image patches and text tokens. This prior allows for more precise alignment, especially for tasks requiring high semantic accuracy.

Weakly Supervised Training: Many alignment approaches rely on large, fully annotated datasets with paired samples for supervised training. However, in our domain, instance-level annotations for multimodal datasets, such as VIS-IR image pairs with corresponding object labels, are either unavailable or highly limited. To address this, we propose a weakly-supervised alignment strategy that generates pseudo-labels using text-to-image similarity, avoiding the need for ground truth annotations.

Bidirectional Alignment: Existing methods predominantly align features in one direction (e.g., from text to image or vice versa). This unidirectional alignment is often insufficient when dealing with complex multimodal interactions where both modalities need to inform each other. Our method introduces bidirectional alignment, ensuring that both text and image features are jointly refined, enhancing the quality of the fusion output.

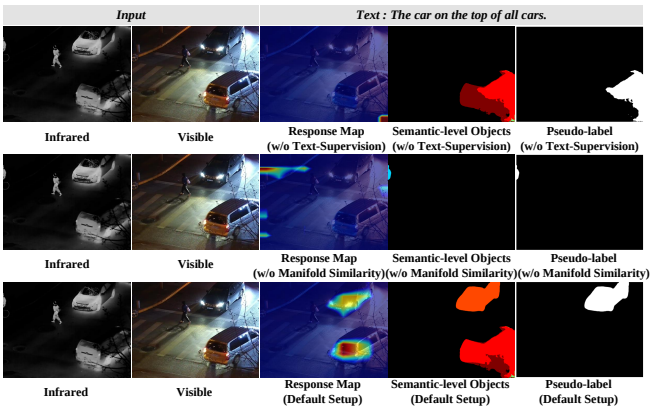

Figure 1. Qualitative results of ablation study on TIRN.

Thus, the unique combination of task-specific instance-level alignment, the utilization of manifold similarity as a geometric prior, weakly-supervised learning, and bidirectional feature alignment distinguishes our method from existing solutions, enabling more effective feature alignment.

4. Comparison with SOTA

Because of space constraints in the main paper, we provide additional qualitative comparison results for various VIS-IR fusion models here, as shown in Fig. 4. It is evident that the uncontrollable image fusion model cannot tailor fused images based on textual input, aligning with the analysis in the Introduction. For the controllable image fusion models, both TextFusion and our model are capable of generating images based on text. However, TextFusion can only highlight multiple objects at the semantic level without precise instance-level localization, whereas our model achieves this level of specificity.

5. Ablation Study

To effectively evaluate the first stage of our model, we construct a VIS-IR dataset that includes text descriptions referring to instance-level objects and binary masks (ground truth, GT) indicating their locations. To obtain these GT masks, we first utilize a large vision foundation model for segmentation [5] to generate preliminary instance segmentation maps. We then select the corresponding instances based on different textual cues and manually refine these maps to ensure accuracy. The numerical results of the ablation experiment, as listed in Table 2 of the main paper, are calculated based on these refined GT maps.

Here, we present qualitative comparison results of ablation experiments in the main paper. As shown in Fig. 1, Fig. 2, and Fig. 3, removing the proposed modules leads to a significant drop in the quality of response maps and pseudo-labels, demonstrating the effectiveness of TIRN, MFA, and the instance selection module.

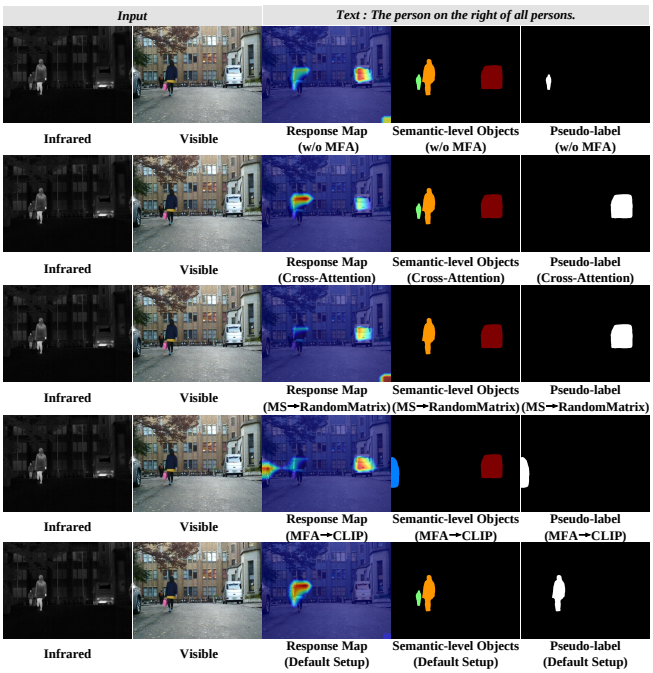

Figure 2. Qualitative results of ablation study on MFA.

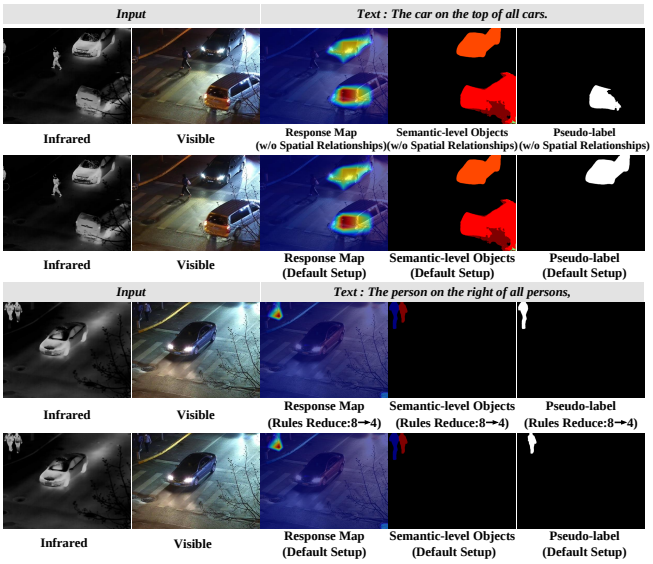

Figure 3. Qualitative results of ablation study on instance selection module.

6. Validation of the Necessity of Tailored Instance Localization for VIS-IR

As discussed in Sections 1 and 2.2 of the main paper, RIS models designed for natural images are unsuitable for object localization in VIS-IR. To verify, we test several RIS mod-

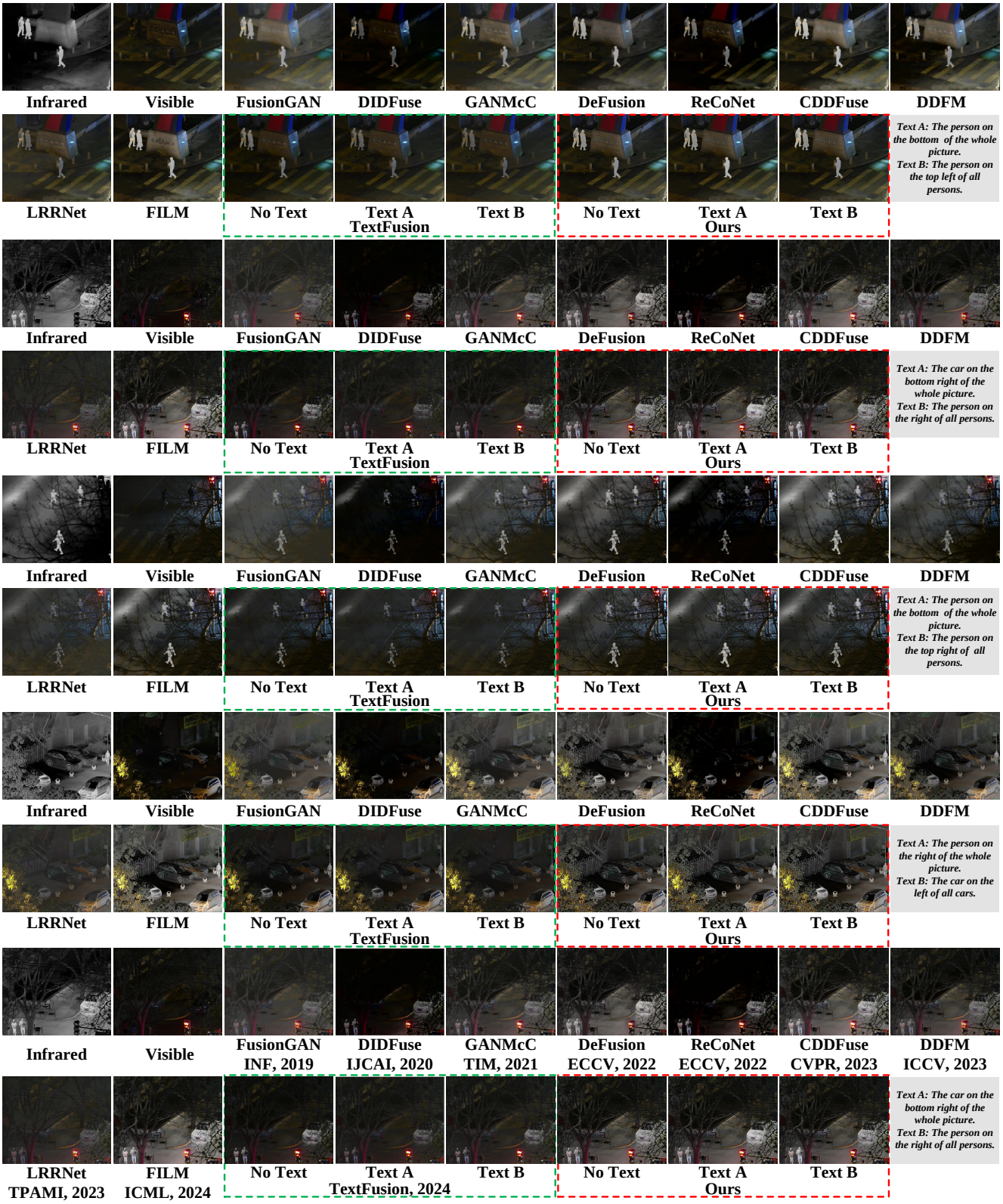

Figure 4. Qualitative comparison of various fusion models. The first nine algorithms are non-controllable fusion models, producing static results regardless of text input. TextFusion and our model are controllable fusion models, displaying fusion results for two different text inputs. TextFusion can only highlight multiple semantic-level objects, whereas our model specifically highlights the referenced instance.

Table 1. Qualitative comparison of our localization method with RIS models on VIS and IR images. Existing RIS models, designed for natural images, can only take either infrared or visible images as single inputs, lacking the ability to integrate complementary information from both modalities for localizing referred objects.

Methods	Precision@0.5	MIoU
LAVT-RIS(Input:IR)	0.000	3.062
LAVT-RIS(Input:VI)	0.787	2.820
VLT-RIS(Input:IR)	0.787	2.401
VLT-RIS(Input:VI)	0.787	1.910
MG-RIS(Input:IR)	25.984	27.811
MG-RIS(Input:VI)	33.070	34.777
Ours(Input:IR and VI)	73.228	66.065

Table 2. Time overhead comparison of various fusion models on two test datasets.

Test Sets	CDDFuse	DDcGAN	EMMA	MetaFusion	MUFusion
TNO	0.0319s	14.8790s	0.3181s	0.0040s	6.9263s
M^3FD	0.0329s	26.5890s	0.4074s	0.0030s	19.2424s
Test Sets	MURF	SDNet	U2Fusion	FILM	Ours
TNO	0.0873s	0.0573s	4.3100s	0.6302s	0.3125s
M^3FD	0.1701s	0.0584s	0.3583s	0.6384s	0.5587s

els (LAVT [9], VLT [2], MG [1]) on IR and VIS images. Unlike these models, which process only single images and yield suboptimal results due to modality limitations, our VIS-IR-tailored method takes two images as input to create a joint visual representation, reducing localization difficulty. Numerical results are listed in Table 1, demonstrating that our model is able to generate more accurate pseudo-labels. Qualitative comparisons are presented in Section 4.5 of the main paper.

7. Runtime Analysis

Table 2 lists the runtime of 10 fusion models. MetaFusion and CDDFuse are faster due to their lower computational complexity. Our model ranks 5th, with an acceptable runtime. The additional processing time is attributed to handling the extra text modality data, which is necessary for our model's functionality. Experiments are conducted on a machine with an RTX 4090 GPU, i9-14900K CPU, 128GB RAM, and PyTorch 2.3.0.

References

[1] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26573–26583, 2024. 5

[2] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7900–7916, 2022. 5

[3] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004. 2

[4] Ron Kimmel, Nir Sochen, and Ravi Malladi. From high energy physics to low level vision. In *Scale-Space Theory in Computer Vision: First International Conference, Scale-Space'97 Utrecht, The Netherlands, July 2–4, 1997 Proceedings 1*, pages 236–247. Springer, 1997. 2

[5] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27948–27959, 2024. 3

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[8] Nir Sochen, Ron Kimmel, and Ravi Malladi. A general framework for low level vision. *IEEE transactions on image processing*, 7(3):310–318, 1998. 2

[9] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 5

[10] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 2