

Ultrasound-Vision Fusion Perception for Visual Failure Compensation in Robotics

Yihan Huang

*School of Mechanical and Power Engineering, East China University of Science and Technology,
Shanghai, China
24011384@mail.ecust.edu.cn*

Abstract. This paper provides a review of robotic ultrasonic–visual fusion covering attempts to solve vision impairments in unstructured environments. While robots have moved beyond structured lab environments to explore diverse environments including warehouses, rescue sites, industrial workshops, and underwater, the reliability of cameras in unstructured environments is highly compromised by reflecting surfaces, variable lighting, and the presence of elements such as smoke, dust, or water. For this reason, a problem arises concerning the way a robot can safely perceive the surrounding unstructured environments, when visual information is either incomplete or gives the wrong impression. Ultrasonic sensing becomes a valuable substitute to visual sensing for the reasons that it does not show a significant level of dependence on surface aesthetics. Robotic usage of ultrasound provides distance or echo based responses to address the information needs created by the degradation of images. The review presents ultrasonic sensing, analyzes major visual failure scenarios, as well as the physical feasibility of compensating vision with ultrasound. It outlines data-level, feature-level, and decision-level fusion methods. For the reasons outlined in the review, mobile robot navigation, robotic manipulation, and underwater exploration are a few of the fields where ultrasonic-visual fusion can help realize robot control tasks. Data heterogeneity, dynamic-scene robustness, calibration, synchronization, and evaluation standards remain open for challenge. Work targeting learning-based representation, lightweight fusion, and realistic deployment benchmarks are warranted.

Keywords: ultrasound-vision fusion, robotic perception, visual failure, multi-sensor information fusion

1. Introduction

Complex robotic operations, including unstructured manipulator grasping and autonomous navigation, rely heavily on perception systems that remain dependable outside controlled laboratory conditions [1]. Vision based techniques tend to work well in the event of adequate illumination and rich texture, as cameras are able to offer both dense space information along with very strong semantic clues. They are, however, less reliable when robots are introduced to real deployment environments with low-light conditions [2]. Backlighting and uneven illumination can further reduce the stability of visual perception by causing local overexposure or loss of image details [3].

In smoke-filled and dust-filled industrial environments, suspended particles can scatter light and seriously degrade the quality of visual information [4]. In recent robotic grasping and navigation research, these failures are not usually due to an inherent weakness in the algorithm. Instead, they often arise from the physical interaction between light and the environment. These blind spots can disrupt the planning of grasps, distort distance calculations, and eventually the number of tasks that a robot can safely perform.

Such a restriction poses a practical question to robotic perception: how can a system be robust as optical sensing is unreliable? A potential solution to the aforementioned issue is to add complementary sensing modalities that are less reliant on visual appearance. Of these, ultrasonic sensing is particularly appealing as it can produce reasonably stationary echo responses in both transparent and specular media, is relatively unaffected by the aerosol in aerosol-rich environments, and can be realized at relatively low cost. However, ultrasound does not automatically enhance things. The actual challenge is presented at fusion. Vision gives dense and high-dimensional image representations, and ultrasound typically gives low-resolution range measurements, or sparse point clouds [5]. Their differences of spatial representation, density of information and signal structure are hard to compensate especially when the robot needs to respond as the visual perception starts to falter.

There has been an increased amount of literature examining ultrasonic-visual perception in local obstacle avoidance and robotic grasping. Even though these studies have shown promising findings, a logical framework has not come out yet. The first issue, which has remained a thorn in the flesh, is that visual failure is almost never determined in a one-fashioned way. A large body of work considers single phenomena, like transparency, reflection, or low-illumination, as opposed to considering failure as just another perception problem with common underlying physical effects and task-level implications. Fusion strategies can be also spread at data-level alignment, feature-level interaction, and decision-level reasoning, and evaluation criteria may heavily rely on particular sensor layouts, experimental platforms, and assumptions on tasks. This disaggregation aids in understanding why promising methods are challenging to compare, replicate, and externalize their results to different test conditions.

It is against this background that a systematic review on the visual failure compensation is needed. In the paper, the author determines the representative failure cases and considers ultrasound-vision fusion structures at various levels of integration. The review will shed light on the interactions of heterogeneous optical and acoustic data, where current methods are useful, and what barriers continue to limit deployment to create a better reference in future research, and a practical foundation in engineering implementation and standardization.

2. Background and current status

The combination of ultrasound and vision has been evolving slowly as it has progressed through the integration of sensors, to more specific forms of compensation of visual deficiency and reveals various underlying issues. Machine vision did not come into robotic systems as much as ultrasonic sensors had. Borenstein and Koren later in the 1980s showed the basic functionality of ultrasound in real time obstacle avoidance [6]. Subsequently, a local spatial sensing scheme using rings of multiple ultrasonic transducers would become a typical setup of mobile robots that do not have a pre-existing map. At this time, ultrasound was not a supplement to vision but rather a separate and often dominant sense organ. Vision technologies were developed at great speed after 2000, however, sudden shifts in illumination, white surfaces, and barriers in real-life settings still led to breakdown

of pure vision-based systems. These recurring failures encouraged researchers to seek compensatory solutions for the physical blind spots of single-modality perception.

2.1. From coarse complementarity to precise compensation for visual failure

As robotic applications have expanded from open indoor floors to tabletop manipulation and harsh media environments, current research no longer treats "complex environments" as a vague and uniform category. Instead, more studies now focus on the concrete boundaries of visual failure, such as depth holes produced by transparent acrylic panels in depth-camera data or feature overexposure in RGB images under strong backlighting. This shift reflects an important change in research practice: compensation is becoming more failure-specific, rather than being designed only as a general robustness enhancement.

The fusion at the feature level is now a major research focus as it allows visual and ultrasonic representations to interact with each other during learning. An example of visual network can create suspicious two-dimensional bounding boxes, which can be checked with directional ultrasonic ranging; or with wide-beam ultrasound, the approximate direction of an obstacle can be estimated and vision enabled with high-precision recognition. Decision-level fusion is still very popular in engineering systems, in which the Bayesian networks and the Dempster-Shafer evidence theory is frequently employed to incorporate the visual class confidence with the ultrasonic distance confidence in a probabilistic form [7].

Simultaneously, the scope of application of fusion technologies is ever-increasing in magnitude and media of operation. Miniature ultrasonic sensors can be used at the manipulator end to give the least distance requirements when grippers are moving towards transparent workpieces on a six-degree-of-freedom robot. The penetrative ability of acoustic sensing also combined with pixel-level optical details in very turbid underwater conditions enables perception systems to store useful spatial information even in case of severe degradation of visual clarity.

2.2. Core challenges and technological gaps

Although fusion is becoming more and more profound, there are still many gaps that can be identified in the existing studies. The first one is the absence of common standards. Current literature varies considerably in both combination algorithms, and hardware design e.g. monocular vision plus single ultrasonic sensor are hard to compare to stereo vision plus ultrasonic array and the extreme paucity of publicly available multimodal datasets similar to KITTI compels research teams to test their approaches in self-assembled setups, which complicate cross-study measurement [8]. The second problem is lack of strength in dynamic and unstructured environments. Most high-precision feature-fusion algorithms are usually designed in a set of quasi-static laboratory tests, including fixed lights and non-portable targets; when applied to real-world workshops with high-frequency vibration, or dynamic pedestrian noise or acoustic multi-path effects, fusion network confidence tends to drop quickly. Lastly, black-box behavior and computational bottlenecks are two challenges encountered by end-to-end architectures. Though the attention mechanisms allow the deep learning models to automatically learn the complementary relationship between ultrasound and vision, which are theoretically beneficial, such dense networks are hard to implement on edge devices with limited resources [9]. Balancing between accuracy, interpretability and lightweight implementation is also critical to the success of this technology shifting out of the lab.

3. Fundamentals of ultrasonic sensing and its applications in robotic perception

Why is it possible to substitute vision with ultrasound? Knowledge about its ranging principle and physical constraints are a precondition to elucidate its engineering role in robot systems.

3.1. Ranging mechanism based on mechanical waves and its inherent advantages

The principle behind ultrasonic sensing refers to the sending and receiving of acoustic waves of very high frequencies and the reflection of the acoustic wave back. The sensor is able to determine real-time spatial data on the local environment by measuring the distance of objects around it and this is achieved by examining the return time of the echo. This process does not require ambient illumination, object color or surface texture as is the case in vision-based perception. Consequently, ultrasonic sensing can continue to be used in conditions of visual degradation like darkness, smoke, dust and low-texture. This property makes it a practical complementary modality for robotic perception, especially when visual input becomes unstable or partially unavailable.

3.2. Physical limitations confine it to a complementary role

However, the physical properties of mechanical waves are a double-edged sword, defining the boundary beyond which ultrasound cannot independently support advanced intelligent tasks: its spatial resolution is extremely low, because the 40 kHz frequency corresponds to a relatively long wavelength and is typically associated with a wide beam angle, so it can provide only a coarse "nearest distance within a fan-shaped region" rather than precise object contours or dense point clouds.

3.3. Two distinct engineering application paradigms

Owing to these extreme strengths and weaknesses, ultrasonic perception in modern robotic engineering has gradually evolved into two contrasting application paradigms: in precise manipulator grasping or indoor AGV navigation, high-dimensional topology is mainly handled by vision. In such cases, ultrasound serves only as a low-cost and reliable one-dimensional distance constraint.

For example, when a robotic gripper approaches a transparent glass cup, an RGB-D camera may fail in ranging because of light transmission. At this moment, the one-dimensional distance measurement from a miniature ultrasonic sensor mounted at the end effector can serve as a mandatory physical safeguard against collision. In mobile robots, it is used to impose veto-like behavioral corrections when vision misses nearby transparent obstacles. In extreme media such as underwater environments, electromagnetic waves undergo exponential attenuation, and beyond only a few meters the surroundings become optically chaotic. In this case, the primary–secondary relationship is completely reversed, and acoustic waves become the only effective carrier of medium- and long-range information.

In both underwater transducer detection and the ultrasonic inspection of dam cracks, ultrasound is responsible for building global spatial awareness. Optical cameras, in contrast, are reduced to auxiliary tools. They only provide local texture information at very close range.

4. Analysis of visual failure scenarios and the feasibility of ultrasonic compensation

The collapse of visual perception is not a random algorithmic error but a physical inevitability determined by underlying optical laws. To build a robust fusion system, it is necessary to first accurately characterize these "optical dead zones" and then exploit the mechanical-wave properties of ultrasound for cross-modal physical compensation.

4.1. Material traps: transmission and specular reflection

The surface material of target objects is a frequent trigger of visual failure.

When facing transparent glass cups or acrylic plates, both visible and infrared light exhibit high transmissivity. In such cases, RGB-D cameras based on structured light or time-of-flight principles often capture the depth of the tabletop behind the glass rather than that of the object itself [10]. This critical misjudgment is a direct cause of physical collisions during robotic grasping.

Under strong workshop lighting, polished metal workpieces produce directional specular reflections. Local overexposure occurs in the image sensor, and the true surface texture is obscured by specular highlights. This can cause visual feature points such as SIFT or ORB to disappear instantly, thereby completely disabling feature-matching-based 3D reconstruction.

The reflection mechanism of ultrasound is independent of photons and relies entirely on differences in acoustic impedance at medium interfaces. Regardless of how transparent the glass is or how reflective the metal is, the acoustic impedance of these solid materials differs from that of air by several orders of magnitude. Therefore, 40 kHz acoustic waves generate strong echoes when striking these materials, enabling the direct output of one-dimensional absolute distance at locations where visual depth sensing breaks down.

4.2. Environmental barriers: photon deficiency and aerosol scattering

Physical changes in the propagation medium can globally degrade the signal-to-noise ratio of the visual sensing chain.

Extreme illumination changes: Passive vision is highly susceptible to local blindness under very low light or backlighting conditions. In contrast, ultrasound operates by actively emitting mechanical vibrations and is therefore completely immune to illumination conditions [2, 3], with its ranging performance unaffected by ambient lux levels. Aerosol-induced blindness: In fire rescue scenarios or dust-filled industrial environments, suspended micrometer-scale particles in the air cause severe optical scattering, reducing LiDAR or vision-based depth maps to irregular noise [4]. The wavelength advantage of ultrasound: The ability of ultrasound to penetrate smoke and dust fundamentally lies in its wavelength. A 40 kHz sound wave has a wavelength on the centimeter scale, which is far larger than micrometer-scale dust particles. As a result, acoustic waves can readily diffract around such particles and maintain a basic level of distance sensing in turbid media where optical sensing is completely blinded.

4.3. Spatial occlusion: physical limits and partial perspective

It must be clarified that ultrasound is not a "see-through-wall radar." When confronted with dense occluders, optical waves and acoustic waves share the same physical limit.

Neither vision nor ultrasound can penetrate concrete walls or thick metal plates, and thus neither can obtain information behind such obstacles. When facing sparse occluders such as wire meshes or

dense branches, pure vision is often severely disturbed by the complex foreground grid. In such cases, the relatively wide beam angle of ultrasound may allow part of the acoustic energy to pass through the gaps in the grid and detect large obstacles behind it. However, this also introduces strong multipath noise. Therefore, the optimal way to handle occlusion is not to rely on ultrasound to penetrate obstacles directly, but to move the robot platform and perform multi-view integration using data from both modalities.

5. Architectures and boundaries of ultrasound–vision cross-modal fusion

Integrating optical and acoustic data involves far more than simply stacking code. Vision produces dense, high-dimensional 2D/3D tensors, whereas ultrasound typically outputs highly sparse one-dimensional range values or wide-beam envelopes. This vast gap in spatial resolution means that the point at which fusion is introduced into the system pipeline directly determines both system robustness and computational cost.

In light of developments in engineering practice, current ultrasound–vision fusion architectures can be clearly divided into three levels [11].

5.1. Forced anchoring, implicit reconstruction, and fail-safe arbitration

This level preserves the most complete information, but it is also the most challenging to implement in engineering practice. Because ultrasound has a relatively wide beam angle, a single echo distance corresponds to an arc-shaped region in space rather than a precise point as in LiDAR. Therefore, forcibly anchoring acoustic data onto high-resolution images can easily lead to spatial ghosting. Once the robot is in high-speed motion, even minor extrinsic calibration errors or timestamp misalignments can cause tearing artifacts at object boundaries in the fused depth map.

The academic community is increasingly shifting toward feature-level fusion, whose key advantage lies in elegantly circumventing the mathematical deadlock of mapping ultrasound's wide beam onto specific pixels. A classic engineering solution is for the vision network to first output a two-dimensional bounding box for a suspicious region, after which ultrasonic features are injected into that feature layer as an absolute physical constraint along the Z-axis. However, such black-box models, which rely heavily on end-to-end training, are often tightly coupled with specific grasping or navigation tasks and are therefore difficult to generalize across platforms.

As the oldest and most reliable solution in industry, decision-level fusion has a very low deployment cost and can even run asynchronously on two independent MCUs. Its logic is uncompromising: once the visual pipeline collapses because of strong light or transparent glass, the ultrasonic pipeline can still independently issue an emergency braking command based on low-level echoes. The trade-off is that the system gives up all fine-grained environmental reconstruction capability.

5.2. Hybrid parallel architectures under computational constraints

Early industrial applications relied heavily on decision-level fusion for robustness, whereas current theoretical research is increasingly focused on deep-learning-driven feature-level fusion. Neural networks are attempting to adaptively learn the nonlinear complementary relationship between visual RGB data and ultrasonic echoes. However, given the stringent power and computational constraints of edge devices such as micro aerial vehicles and robotic end-effector controllers, future general-purpose perception systems are unlikely to rely on a single fusion level alone. A feasible

path is a hybrid parallel architecture, in which lightweight decision-level fusion remains continuously active to guarantee collision safety, while computationally expensive feature-level fusion is activated on demand during fine operations.

6. Engineering gaps and directions of technological evolution

Current ultrasound–vision fusion systems mostly remain at the proof-of-concept stage. To move from static laboratories to real unstructured industrial scenarios, the field must overcome the following three major engineering gaps. Future technological evolution will also revolve closely around these pain points.

6.1. Representation gap: from naive concatenation to cross-modal implicit alignment

The representation gap is the fundamental bottleneck of heterogeneous fusion. Vision outputs high-dimensional dense matrices, whereas ultrasound outputs low-dimensional sparse scalars or wide-beam envelopes. Early studies attempted to combine these two data types through simple channel concatenation or manually designed rules, but the acoustic envelope details of ultrasound were often lost during feature extraction. Although more recent end-to-end networks can learn implicit cross-modal alignment, they rely heavily on large amounts of high-quality aligned data, and their black-box nature still faces a trust crisis in safety-critical industrial control [12].

6.2. Deployment barriers: calibration drift and edge-computing constraints

A key practical bottleneck is that high-precision feature-level fusion relies on absolutely rigid extrinsic calibration between visual cameras and ultrasonic transducers. However, during the actual operation of quadruped robots or micro aerial vehicles, high-frequency vibration and abrupt temperature changes can cause submillimeter mechanical deformation of the sensors. When amplified onto a sensing surface several meters away, such minute calibration drift can lead to severe tearing in multimodal projection.

On the one hand, research is shifting from offline static calibration toward online adaptive calibration based on environmental features. On the other hand, complex drift-resistant networks and multimodal Transformer models directly encounter the computational limitations of edge devices. How to aggressively prune and streamline fusion algorithms on computation-constrained embedded platforms is a critical factor determining whether the technology can be practically deployed [13].

6.3. Dynamic noise and the lack of benchmarks

The high accuracy reported in much of the existing literature is often achieved on static tabletops or in empty corridors. Once deployed in real workshops, dynamic pedestrians and moving AGVs can frequently disrupt visual tracking, while complex metallic mechanical structures can generate severe acoustic multipath effects, causing ultrasound to detect numerous false-positive obstacles. The greatest obstacle to breakthroughs in this field is the absence of a unified public multimodal dataset comparable to KITTI or TUM in vision-based autonomous driving. Different laboratories use ultrasonic sensors with different frequencies and cameras with different fields of view, and they often report results only in their own self-built scenarios. A primary task for the future is to establish standardized testbeds and stress-test benchmarks for extreme edge scenarios; otherwise, fair cross-comparison among different algorithms will remain impossible.

7. Conclusion

For robots to function reliably beyond structured, static workshops, perception must overcome the blind zones that inevitably arise from any single sensing modality. This review suggests that ultrasound-based visual compensation should not be understood as a simple stacking of sensors and algorithms. Rather, it reflects a physically grounded route toward stronger perceptual robustness. This point becomes especially clear in recent robotic studies involving transparent containers, polished metal parts, dim corridors, or dust-filled industrial scenes, where camera performance often drops sharply despite otherwise strong visual models. Because ultrasound relies on acoustic impedance contrast and long-wavelength diffraction, it can still return useful cues in situations where optical sensing is inherently fragile, including transparent materials, highly reflective surfaces, low-visibility media, and environments disturbed by smoke, dust, or turbidity. Under such conditions, photons may be lost through transmission, overexposure, or scattering, while acoustic signals can still contribute complementary distance or boundary information.

The main obstacle lies in the mismatch between the two modalities themselves. Vision offers dense, high-dimensional spatial and semantic descriptions, whereas ultrasound more often produces sparse one-dimensional ranging data, coarse profiles, or wide-beam echo responses. In practical system design, this gap means that ultrasound–vision fusion cannot be treated as a plug-and-play framework. The fusion strategy has to be shaped by task demands, sensing geometry, available computation, and environmental disturbance. For mobile robots, ultrasound often serves best as reinforcement for near-field safety and collision avoidance. In manipulator systems, it can supply minimum-distance constraints during the grasping of transparent or reflective objects. Underwater, or in strongly turbid environments, the balance may shift further, with acoustic sensing becoming primary and vision contributing local texture or semantic detail only when conditions permit.

Seen from this perspective, future progress should not be measured only by higher detection or recognition scores. A more pressing need is to reduce the fragmentation that still characterizes evaluation practice by building standardized multimodal benchmarks for dynamic, unstructured, and strongly disturbed environments. Without such benchmarks, it remains difficult to judge whether a reported fusion method truly improves robustness or simply adapts to a particular dataset, sensor arrangement, or laboratory setting. At the same time, the field must engage more directly with the physical costs of ultrasound itself, especially its limited spatial resolution, broad beam pattern, and vulnerability to multipath interference. A mature ultrasound–vision framework will therefore need to balance the representational strength of end-to-end implicit reconstruction against the efficiency, interpretability, and reliability of lightweight logical arbitration. Only through that balance can ultrasound–vision fusion move beyond task-specific demonstrations and become a dependable perceptual basis for robots operating in genuinely unstructured environments.

References

- [1] Mao, J., Shi, S., Wang, X., & Li, H. (2023). 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131, 1909–1963. <https://doi.org/10.1007/s11263-023-01790-1>
- [2] Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., & Zhang, Y. (2023). Retinexformer: One-stage Retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 12504–12513). <https://doi.org/10.1109/ICCV51070.2023.01150>
- [3] Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.-M., Gu, J., & Loy, C. C. (2022). Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12), 9396–9416. <https://doi.org/10.1109/TPAMI.2021.3126387>

- [4] Song, Y., He, Z., Qian, H., & Du, X. (2023). Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32, 1927–1941. <https://doi.org/10.1109/TIP.2023.3256763>
- [5] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Zhang, Z., Dong, C., & Tao, D. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 87–110. <https://doi.org/10.1109/TPAMI.2022.3152247>
- [6] Borenstein, J., & Koren, Y. (1989). Real-time obstacle avoidance for fast mobile robots. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5), 1179–1187. <https://doi.org/10.1109/21.44033>
- [7] Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), 447–458. <https://doi.org/10.1109/34.55104>
- [8] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3354–3361). <https://doi.org/10.1109/CVPR.2012.6248074>
- [9] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4015–4026). <https://doi.org/10.1109/ICCV51070.2023.00371>
- [10] Sajjan, S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., & Song, S. (2020). ClearGrasp: 3D shape estimation of transparent objects for manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3634–3642). <https://doi.org/10.1109/ICRA40945.2020.9197518>
- [11] Wang, Y., Mao, Q., Zhu, H., Ma, X., Wang, Y., Deng, Z., Zhang, C., Ji, J., & Li, Y. (2023). Multi-modal 3D object detection in autonomous driving: A survey. *International Journal of Computer Vision*, 131, 2122–2152. <https://doi.org/10.1007/s11263-023-01784-z>
- [12] Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., & Li, H. (2023). Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 17853–17862). <https://doi.org/10.1109/CVPR52729.2023.01712>
- [13] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7464–7475). <https://doi.org/10.1109/CVPR52729.2023.00721>