MAGIC: LEARNING FROM VISIBILITY ASYMMETRY FOR UNSUPERVISED STEREO MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning disparity in occluded regions remains difficult for unsupervised stereo matching. Photometric supervision lacks valid target-view correspondences in these regions, while the teacher and student in conventional binocular self-training share the same target view and therefore the same occlusions. Even when supervision is available, the small proportion of occluded pixels limits their contribution to training. We propose **MAGIC**, a multi-baseline geometric consistency framework for reliable and effective occlusion supervision. The teacher and student share a reference image but use different target views, allowing the teacher to observe correspondences that are occluded from the student. After aligning disparities across baselines, MAGIC uses predictions from teacher-visible regions to supervise student-occluded regions. An occlusion-aware weighting strategy strengthens supervision on teacher-visible but student-occluded pixels, preventing their training signal from being overwhelmed by non-occluded regions. We also introduce **MBS20K**, a synthetic multi-baseline stereo dataset spanning diverse scenes, weather, and lighting. Pre-trained on MBS20K, MAGIC generalizes to real-world datasets with consistently fewer occluded-region outliers. On KITTI, the pre-trained model already outperforms several fine-tuned unsupervised methods. Fine-tuning this model on standard binocular pairs achieves state-of-the-art unsupervised performance on KITTI 2015 and 2012. Incorporating synthesized multi-baseline views during fine-tuning further improves performance. Our code and dataset will be released upon acceptance.

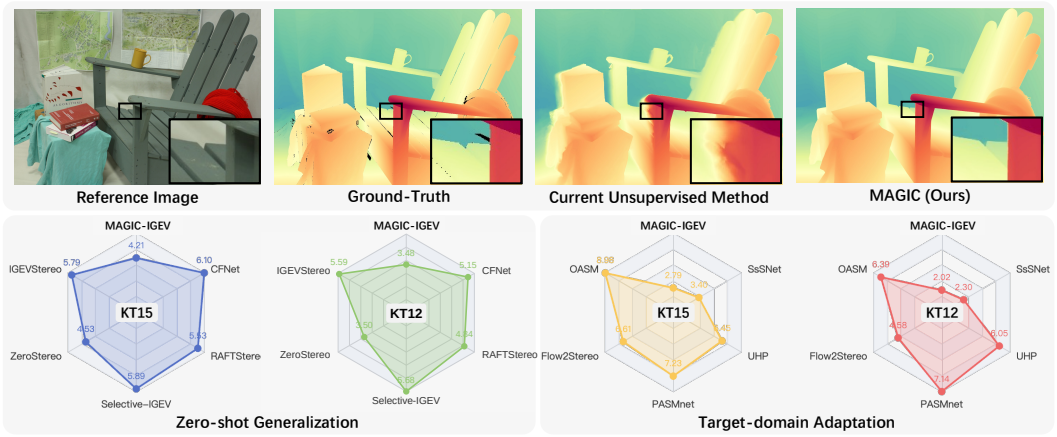


Figure 1: **Top:** MAGIC produces accurate disparity estimates in occluded regions (black box; please zoom in for more details). **Bottom:** MAGIC achieves outstanding performance in both zero-shot generalization (left) and target-domain adaptation (right) on KITTI 2015 and 2012 (Menze & Geiger, 2015; Geiger et al., 2012) (lower is better).

1 INTRODUCTION

Stereo matching recovers dense correspondences between rectified image pairs and serves as a fundamental component in applications such as autonomous driving, robotics, and augmented reality. As ground-truth disparities are costly to acquire, unsupervised methods have attracted increasing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

attention by learning directly from unlabeled image pairs, most commonly through photometric consistency between the reference and target views (Zhong et al., 2017).

This correspondence-based supervision breaks down in occluded regions[1]. Without a valid correspondence in the target image, disparity must be inferred from contextual cues, making its estimation a *completion* task. Applying the photometric loss in these regions misleads learning (Fig. 1). A common remedy detects occlusions and excludes them from the loss (Li & Yuan, 2018). Although this suppresses erroneous gradients, it leaves disparity completion under-constrained.

Pseudo-label self-training (Pang et al., 2018) provides an alternative source of supervision but shares the same visibility limitation under a standard binocular setup. A teacher network predicts pseudo-disparities for the student, yet both networks observe the same target view. Regions occluded to the student therefore remain invisible to the teacher. The teacher's pseudo-labels can be unreliable in occluded regions, and their errors can propagate to the student during training.

A second challenge remains even when reliable occlusion supervision is available. Occluded pixels constitute only a small fraction of an image, so their learning signal can be overwhelmed by the abundant non-occluded regions. Effective occlusion learning therefore requires not only reliable supervision, but also a mechanism that prevents such supervision from being diluted (Fig. 4).

To obtain reliable supervision in occluded regions, we revisit a common assumption in self-training that the teacher and student must observe the same target view. When additional viewpoints with varying baselines are available, a region occluded in one view is frequently visible in another. Moreover, as long as both networks share the same reference image, the underlying scene geometry remains consistent, so disparities predicted under different baselines can be aligned through a simple scaling relation. This visibility asymmetry enables the teacher to rely on valid correspondence cues in its own view and to provide reliable pseudo-labels for regions that are occluded to the student.

Building on this observation, we propose MAGIC, a multi-baseline geometric consistency framework for unsupervised stereo matching. MAGIC retains the teacher-student paradigm (Caron et al., 2021) but assigns distinct target views to the two networks, and rescales the teacher's disparities by the baseline ratio to obtain pseudo-labels that satisfy geometric consistency for the student. To address the supervision imbalance, we further introduce an occlusion-aware weighting strategy that detects occluded regions in the teacher and student views, then amplifies supervision on teacher-visible yet student-occluded pixels and discards the teacher's unreliable pseudo-labels.

To support the training of MAGIC, we construct MBS20K, a synthetic multi-baseline stereo dataset rendered in the CARLA simulator (Dosovitskiy et al., 2017) using a five-camera rig under diverse weather and illumination conditions. Its 20K synchronized frame sets can be combinatorially expanded into millions of training triplets.

Pre-trained on MBS20K, MAGIC exhibits strong zero-shot generalization across diverse real-world scenes and weather conditions, reducing occluded-region outliers on KITTI 2015 and 2012 by 27.3% and 36.2%, respectively. When submitted to the KITTI benchmarks without any fine-tuning, it already outperforms several fine-tuned unsupervised methods. After fine-tuning with only standard binocular pairs, MAGIC establishes a new unsupervised state-of-the-art, owing to its robust pre-training, and improves further when multi-baseline observations are available in the target domain. Our contributions are summarized as follows:

- We propose MAGIC, a teacher–student framework that exploits visibility asymmetry and transfers disparity supervision across different baselines through baseline-ratio rescaling.
- We identify supervision imbalance as a key factor hindering learning in occluded regions and address it with an occlusion-aware weighting strategy that upweights sparse occluded pixels to learn accurate disparity completion in these regions.
- We construct MBS20K, a synthetic multi-baseline stereo dataset for pre-training with geometric consistency.
- In the zero-shot setting, MAGIC generalizes strongly; after binocular fine-tuning, it establishes a new unsupervised state-of-the-art on KITTI 2015 and 2012, with further gains from multi-baseline target-domain data.

[1]We treat a reference pixel as *occluded* if its corresponding scene point is hidden behind another surface in the target view or lies outside the target camera's field of view.

2 Related Work

Deep stereo matching. Stereo matching has evolved over the past decades from hand-crafted pipelines (Scharstein & Szeliski, 2002) to end-to-end deep learning frameworks (Mayer et al., 2016). Modern deep stereo networks generally fall into two architectural families: 3D convolution-based cost volume regularization (Kendall et al., 2017; Guo et al., 2019; Shen et al., 2021) and ConvGRU-based iterative refinement (Lipson et al., 2021; Li et al., 2022; Zhao et al., 2023). IGEVStereo (Xu et al., 2023) employs a lightweight 3D convolutional module to provide strong geometric initialization for iterative refinement, and Selective-Stereo (Wang et al., 2024) aggregates multi-frequency representations to improve performance in edge and textureless regions. More recent approaches (Jiang et al., 2025; Cheng et al., 2025; Wen et al., 2025; 2026) further incorporate monocular depth priors derived from foundation models (Yang et al., 2024b).

Photometric-based supervision. Photometric consistency provides a natural supervisory signal for training stereo networks on unlabeled data (Zhong et al., 2017). To alleviate photometric failure in occluded regions, OASM (Li & Yuan, 2018) masks photometrically inconsistent pixels, PASM-net (Wang et al., 2022) suppresses unreliable regions via attention-based confidence filtering, and UHP (Yang et al., 2024c) incorporates planar geometric priors. However, masking occluded areas removes their direct supervision, leaving disparity completion under-constrained.

Pseudo-label-based self-training. Self-training supervises students with teacher pseudo-disparities. ZOLE (Pang et al., 2018) exploits zoomed-in patches for additional image detail, and DualNet (Wang et al., 2025b) uses feature-metric and augmentation consistency to train a teacher, then freezes it for student supervision. Other methods update teachers via an exponential moving average (EMA) (Grill et al., 2020). Zhang et al. (2024) combine frozen and EMA teachers with ground truth and pseudo-labels to preserve generalization during stereo fine-tuning. TS-PSMNet (Yue et al., 2024) filters uncertain predictions by confidence, while CST-Stereo (Zhou et al., 2025) extracts reliable pseudo-labels through multi-resolution and iterative consistency. With the same target viewpoint, however, both networks share the same occlusions, leaving the teacher without additional correspondence cues where the student is occluded.

Multi-view supervision. NeRFStereo (Tosi et al., 2023) uses a central reference and two target views, taking the per-pixel minimum reconstruction error to alleviate occlusion. SMV (Yuan et al., 2021) aligns center-left and center-right disparities through uncertainty gating, allowing the confident view to correct the other in occluded regions. Both methods use equal baselines. In self-supervised multi-view stereo (MVS), KD-MVS (Ding et al., 2022) validates depths across views before probabilistic teacher–student distillation. For stereo disparity transfer, MAGIC rescales teacher predictions by the baseline ratio, removing the equal-baseline constraint and combinatorially exploiting multi-baseline data. Moreover, MAGIC applies occlusion-aware weighting to strengthen supervision in occluded regions, whose sparse signal is otherwise overwhelmed by supervision from the abundant non-occluded pixels throughout the training images.

3 Method

3.1 Preliminary

Given a pair of rectified stereo images, stereo matching estimates, for each pixel in the reference image I_{ref}, its disparity d to the corresponding pixel in the target image I_{tgt}. The depth is recovered by triangulation as $z = B \cdot f/d$, where B is the baseline and f is the focal length.

Unsupervised methods commonly supervise this prediction through photometric consistency (Zhong et al., 2017). The target image is warped by the predicted disparity to reconstruct the reference view $\hat{I}_{ref}$, and a photometric loss combines SSIM and L_1 terms,

$$\mathcal{L}_p = \frac{\alpha}{2}(1 - \text{SSIM}(I_{ref}, \hat{I}_{ref})) + (1 - \alpha)||I_{ref} - \hat{I}_{ref}||_1. \tag{1}$$

To complement photometric supervision, an edge-aware smoothness term $\mathcal{L}_s$ encourages locally smooth disparities while preserving discontinuities at image edges (Zhong et al., 2017).

Another line of unsupervised methods uses a teacher to supervise a student with pseudo-disparities (Pang et al., 2018). Throughout, superscripts s and t denote the student and teacher,

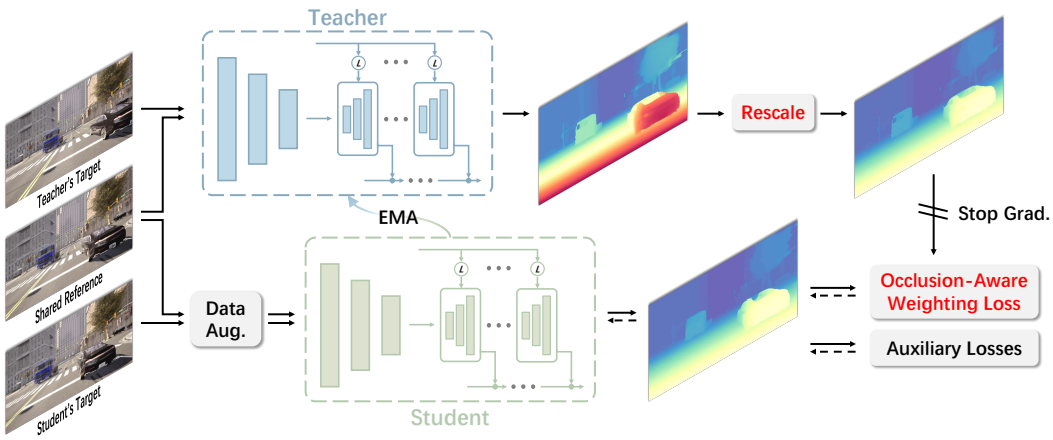


Figure 2: **Overview of MAGIC.** The illustrated input triplet shares a reference image but uses different target views, letting the teacher observe regions occluded in the student's view. Teacher disparities are rescaled by the baseline ratio to match the student's disparity scale. The resulting pseudo-labels are detached from the computational graph and used to supervise the student. Occlusion-aware weighting strengthens supervision on sparse occluded pixels so that their contribution is not overwhelmed by abundant non-occluded pixels. Data augmentation is applied only to the student's input, while the teacher is updated as an EMA of the student's weights.

respectively. Given their predicted disparities d^s and d^t, the student is supervised by $|d^s - \mathrm{sg}(d^t)|$, where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator applied to the teacher prediction.

However, neither approach provides reliable supervision in occluded regions. Photometric supervision lacks valid correspondences in these regions. In binocular self-training, the teacher also lacks matching evidence and must solve the more challenging task of disparity completion, potentially producing unreliable pseudo-labels that propagate errors to the student.

3.2 Multi-Baseline Geometric Consistency

MAGIC is an unsupervised framework that exploits geometric consistency across stereo observations captured with different reference-to-target camera baselines.

Asymmetric input viewpoints. Based on the teacher-student paradigm, MAGIC introduces visibility asymmetry. As illustrated in Fig. 2, the student and teacher networks share the same reference image I_{ref} but receive independently sampled target views I^s_{tgt} and I^t_{tgt}. Different target views yield complementary visibility patterns. Pixels occluded in the student's target view may remain visible in the teacher's view. Relying on valid correspondence cues available in its own view, the teacher can predict accurate disparities for these student-occluded regions.

Since both networks operate on the same reference image, the underlying 3D scene geometry remains consistent between them. For a pixel with depth z in the reference image, the predicted disparities d^s and d^t are related by

$$\frac{B^s \cdot f}{d^s} = \frac{B^t \cdot f}{d^t} = z \quad \Rightarrow \quad d^s = r \cdot d^t,\ r = \frac{B^s}{B^t}. \tag{2}$$

Rescaling the teacher's disparity by the baseline ratio r maps the teacher's prediction into the student's disparity space, enabling supervision to be transferred across baselines.

Robust student and momentum teacher. MAGIC also introduces appearance asymmetry through student-only augmentation, including color jitter and random occlusion. These more challenging student inputs encourage robust features for disparity estimation (Yang et al., 2024a).

The teacher is initialized from the student weights and updated using an EMA of the student's weights during training. The update rule is $\theta^t \leftarrow m \cdot \theta^t + (1 - m) \cdot \theta^s$, where θ denotes the model weights, and m is a momentum parameter following a cosine schedule (Grill et al., 2020). Empirically, a momentum teacher consistently improves performance compared to a fixed teacher. During inference, only the teacher is used for disparity estimation.

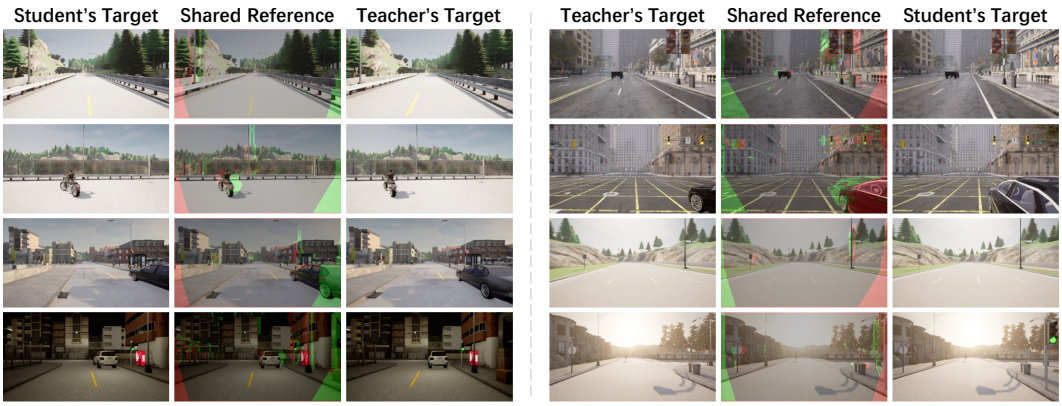


Figure 3: **Visualization of the occlusion-aware weight map.** We show two triplet arrangements. Colors indicate the visibility regions defined in Section 3.3. Red areas denote *teacher-occluded* regions, where supervision is discarded due to unreliable teacher predictions. Green areas denote *student-occluded yet teacher-visible* regions, where supervision is emphasized. The remaining areas denote *co-visible* regions. To highlight the visibility differences between the teacher and student views, auto-masking (Godard et al., 2019) is disabled here.

Geometric consistency loss. The primary supervision signal enforces consistency between the student's prediction and the rescaled teacher disparity:

$$\mathcal{L}_g = \left\| d^s - r \cdot \mathrm{sg}(d^t) \right\|_1 . \tag{3}$$

Unlike binocular self-training, this geometric consistency objective reliably supervises student-occluded regions by transferring geometrically aligned predictions from teacher-visible regions.

3.3 Occlusion-Aware Weighting

For distinct target views, multi-baseline transfer involves three visibility regions in the reference image: (1) *co-visible* regions, where both the student and teacher observe valid correspondences; (2) *teacher-occluded* regions, where the teacher lacks reliable correspondence cues; and (3) *student-occluded yet teacher-visible* regions, where supervision for the student's disparity completion becomes uniquely available through visibility asymmetry.

Region (2) introduces unreliable pseudo-labels, whereas region (3) provides teacher guidance where student correspondences fail. Appendix E provides an analysis of the effectiveness of teacher pseudo-labels in region (3). However, obtaining this supervision does not ensure that it contributes effectively to learning. Neural networks tend to prioritize simple and frequent patterns (Arpit et al., 2017), so uniform weighting allows abundant co-visible pixels to dominate optimization, limiting disparity completion learning in the sparse region (3).

To address this supervision imbalance and filter out erroneous guidance, we introduce an occlusion-aware weight map $\mathcal{A}$, defined over the computed teacher and student visibility masks $\mathcal{V}^t, \mathcal{V}^s \in \{0, 1\}$, where $\mathcal{V} = 1$ indicates a visible pixel. Each pixel is then assigned a weight as

$$\mathcal{A} = \mathcal{V}^t \big(\mathcal{V}^s + \omega \, (1 - \mathcal{V}^s) \big), \tag{4}$$

where unreliable supervision signals from region (2) are explicitly discarded, and a hyperparameter $\omega > 1$ increases the penalty assigned to region (3). This weighting encourages the student to learn disparity completion from reliable teacher predictions in region (3).

The visibility masks $\mathcal{V}^t$ and $\mathcal{V}^s$ are computed using two complementary criteria. Since occluded pixels typically incur large photometric errors, we threshold the teacher and student photometric losses, $\mathcal{L}_p^t$ and $\mathcal{L}_p^s$, at τ to identify likely occlusions in their respective target views. In addition, we adopt the auto-masking strategy (Godard et al., 2019) to filter out unreliable stationary pixels (in stereo matching, these correspond to infinite-depth regions such as the sky). A pixel is retained only when it passes both criteria. Note that $\mathcal{L}_p^t$ is used only for mask computation and does not update the teacher. Figure 3 visualizes the occlusion-aware weight maps for the two triplet arrangements, and Appendix E provides an evaluation of the reliability of the computed visibility masks.

Table 1: **Zero-shot generalization across diverse scenes.** We report Out-x in occluded (OCC), non-occluded (NOC), and all (ALL) regions on four real-world datasets. † and ‡ denote methods that incorporate real data or foundation priors, respectively.

Method	KITTI 2015			KITTI 2012			Middlebury			ETH3D		
	OCC	NOC	ALL	OCC	NOC	ALL	OCC	NOC	ALL	OCC	NOC	ALL
PSMNet (2018)	47.64	28.13	28.42	63.20	26.50	27.32	62.30	30.18	34.51	28.56	14.74	15.39
GwcNet (2019)	29.07	12.17	12.53	45.65	11.91	12.67	47.13	20.41	24.11	21.37	10.49	11.09
CFNet (2021)	16.42	5.87	6.10	30.25	4.58	5.15	44.55	16.33	20.22	11.89	5.57	5.87
RAFTStereo (2021)	12.70	5.34	5.53	28.35	4.29	4.84	28.00	9.06	11.96	6.02	2.85	3.04
Selective-IGEV (2024)	13.82	5.70	5.89	31.85	5.06	5.68	22.59	6.73	9.17	9.81	4.07	4.43
NeRFStereo† (2023)	14.62	5.23	5.43	26.97	3.51	4.04	31.10	6.79	10.38	8.35	2.78	3.07
ZeroStereo-IGEV† (2025a)	10.53	4.36	4.53	19.41	3.16	3.50	21.27	4.90	7.40	6.25	1.93	2.19
DEFOMStereo‡ (2025)	12.57	4.79	4.99	21.95	3.83	4.21	20.64	4.39	6.91	5.14	2.08	2.24
Fast-FoundationStereo‡ (2026)	10.18	3.39	3.57	22.10	2.78	3.19	20.41	2.33	5.05	11.26	1.21	1.72
IGEVStereo (2023)	14.26	5.60	5.79	33.66	4.92	5.59	24.28	7.25	9.91	9.76	4.06	4.39
MonSter‡ (2025)	9.64	3.30	3.45	18.82	3.03	3.37	18.42	5.86	7.70	3.53	1.32	1.45
FoundationStereo‡ (2025)	8.10	3.07	3.20	15.28	2.63	2.91	8.53	1.12	2.26	2.65	0.52	0.61
MAGIC-IGEV	10.37	4.05	4.21	21.49	3.08	3.48	23.46	6.34	8.94	7.98	2.51	2.76
MAGIC-MonSter‡	8.17	**2.53**	**2.68**	17.26	**2.16**	**2.48**	17.78	5.59	7.13	3.47	1.19	1.35
MAGIC-FoundationStereo‡	**7.89**	2.61	2.75	**13.57**	2.26	2.50	**8.43**	**1.11**	**2.24**	**2.58**	**0.51**	**0.60**

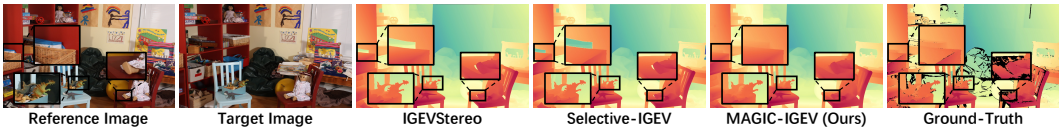


Figure 4: **Qualitative comparison in occluded areas.** Despite being trained with full supervision, the competing methods still produce erroneous disparities in occluded regions. In contrast, by providing reliable and up-weighted pseudo-supervision, MAGIC improves disparity completion.

3.4 Final Loss

Optimizing only the baseline-scaled geometric consistency loss $\mathcal{L}_g$ may admit degenerate solutions (*e.g.*, predicting zero disparity, which trivially satisfies the loss without capturing scene structure). To stabilize optimization and encourage meaningful correspondence learning, we additionally apply the photometric loss $\mathcal{L}_p^s$ and the smoothness loss $\mathcal{L}_s^s$ to the student branch, promoting appearance-based correspondence matching and continuity of local disparity predictions.

The final loss is formulated as

$$\mathcal{L} = \mathcal{L}_g \odot \mathcal{A} + \lambda_p \mathcal{L}_p^s \odot \mathcal{V}^s + \lambda_s \mathcal{L}_s^s, \tag{5}$$

where $\odot$ denotes element-wise multiplication. The occlusion-aware weight map $\mathcal{A}$ balances supervision across different visibility regions, while $\mathcal{V}^s$ masks unreliable photometric supervision in the student-occluded regions. Following the loss configuration of (Godard et al., 2019), we set the weighting parameters to $\alpha = 0.85$ in Eq. (1) and $\lambda_s = 0.001\lambda_p$ in Eq. (5).

These objectives are complementary: photometric loss establishes basic matching, while teacher supervision progressively improves completion where the student's photometric assumption fails. EMA gradually transfers this improved completion ability back to the teacher. Appendix D presents experimental results showing that our framework can be trained from random initialization.

3.5 Training Data

For large-scale unsupervised training, we construct a synthetic multi-baseline stereo dataset, MBS20K, using the CARLA simulator (Dosovitskiy et al., 2017).

Dataset overview. At each timestamp, a rig of five horizontally aligned RGB cameras captures synchronized images, with a uniform baseline of 0.5 m between adjacent cameras. Each camera has a focal length of 480 px and renders images at a resolution of 540×960 px. In total, 20K multi-baseline image sets are collected across diverse scenes under varying weather and illumination conditions. Appendix B provides further dataset details.

Table 2: **Zero-shot generalization across weather conditions** on DrivingStereo. † and ‡ denote methods that use real training data or incorporate foundation-model priors, respectively.

Method	Sunny	Cloudy	Rainy	Foggy
RAFTStereo (2021)	4.23	4.19	12.77	3.09
Selective-IGEV (2024)	5.05	5.24	13.51	4.10
NeRFStereo† (2023)	2.88	2.91	10.20	3.93
ZeroStereo-IGEV† (2025a)	2.59	2.36	13.07	1.89
DEFOMStereo‡ (2025)	3.61	3.75	13.53	2.88
Fast-FoundationStereo‡ (2026)	7.08	6.96	32.65	9.83
IGEVStereo (2023)	4.59	5.15	15.47	4.49
MonSter‡ (2025)	3.48	3.17	5.27	5.03
FoundationStereo‡ (2025)	3.97	2.73	28.50	3.11
MAGIC-IGEV	2.15	1.87	5.07	**1.64**
MAGIC-MonSter‡	**1.91**	**1.78**	**2.35**	2.11
MAGIC-FoundationStereo‡	2.20	2.27	4.18	2.18

Table 3: **Results on KITTI benchmarks**. The unmarked MAGIC-IGEV denotes no fine-tuning, * denotes binocular fine-tuning, and § denotes multi-baseline fine-tuning.

Method	KITTI 2015 (D1)		KITTI 2012 (Out-3)	
	NOC	ALL	NOC	ALL
SsSnet (2017)	3.06	3.40	2.30	3.00
OASM (2018)	7.39	8.98	6.39	8.60
Flow2Stereo (2020)	6.29	6.61	4.58	5.11
PASMnet (2022)	6.69	7.23	7.14	8.57
UHP (2024c)	5.93	6.45	6.05	7.09
DualNet (2025b)	2.67	2.92	2.06	2.59
MAGIC-IGEV	4.27	4.60	2.72	3.27
MAGIC-IGEV*	2.59	2.79	2.02	2.36
MAGIC-IGEV§	**2.45**	**2.65**	**1.92**	**2.31**

Sampling strategy. During training, one image is randomly selected from the five views as the reference image, while two target views are independently sampled with replacement from the remaining four images. Unlike conventional binocular stereo setups, target views are not restricted to the right side of the reference image. If a selected target view lies on the left, the reference and that target are horizontally flipped before being fed into the network, and the predicted disparity is flipped back afterward. Through this sampling strategy, the 20K multi-baseline image sets can be expanded into 1.6M distinct triplet training samples.

4 Experiments

4.1 Experimental Setup

Evaluation protocols. (1) We pre-train MAGIC on MBS20K without disparity labels and evaluate its zero-shot generalization on five public real-world datasets: KITTI 2012 (Geiger et al., 2012), KITTI 2015 (Menze & Geiger, 2015), Middlebury (H) (Scharstein et al., 2014), ETH3D (Schops et al., 2017), and DrivingStereo (Yang et al., 2019). KITTI, Middlebury, and ETH3D provide masks for quantitative occluded-region evaluation. (2) For a fair comparison with existing unsupervised methods, we further fine-tune the pre-trained model on KITTI 2012 and 2015 using standard binocular image pairs. Predictions on the test sets are submitted to the online benchmarks. (3) To evaluate multi-baseline target-domain adaptation, we generate additional viewpoints for the KITTI training data to construct a multi-baseline fine-tuning setting.

Evaluation metrics. We employ Out-x, which measures the percentage of pixels whose absolute disparity error exceeds x px. Following standard practice, we report Out-3 on KITTI and DrivingStereo, Out-2 on Middlebury, and Out-1 on ETH3D. Additionally, we report the D1 error metric for evaluation on the KITTI 2015 benchmark (Menze & Geiger, 2015).

Implementation details. We conduct experiments with IGEVStereo (Xu et al., 2023), MonSter (Cheng et al., 2025), and FoundationStereo (Wen et al., 2025). Following (Wang et al., 2025b), all networks are initialized with official weights trained on SceneFlow (Mayer et al., 2016). For pre-training, models are trained on MBS20K for 200K iterations with a batch size of 16. We use Adam (Kingma & Ba, 2015) with a one-cycle learning rate schedule (Smith, 2018) and a maximum learning rate of 2×10^{-4}. For fine-tuning, models are trained for 50K iterations with the maximum learning rate reduced to 5×10^{-5}. We set $\lambda_p = 10$, $\tau = 0.1$, and $\omega = 2$. Appendix F.1 documents the grid search used to select these hyperparameters.

4.2 Zero-Shot Generalization

To assess the generalization of models pre-trained on synthetic data to unseen real-world environments, we evaluate zero-shot performance across different scenes and weather conditions.

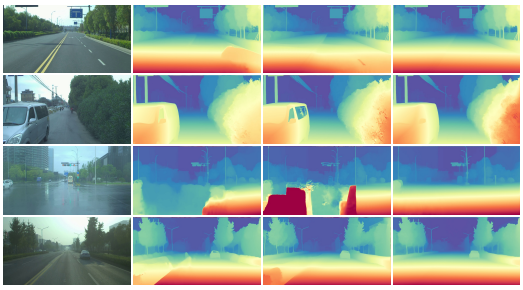


Figure 5: **Qualitative comparison** on DrivingStereo (Yang et al., 2019). From top to bottom: sunny, cloudy, rainy, and foggy.

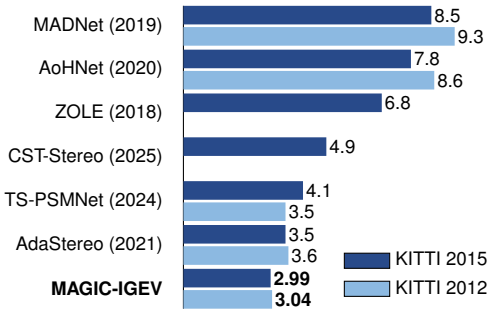


Figure 6: **Results** on KITTI training sets. ZOLE (Pang et al., 2018) and CST-Stereo (Zhou et al., 2025) do not report KITTI 2012 results.

Cross-scene evaluation. As shown in Table 1, MAGIC-IGEV exhibits robust zero-shot generalization, consistently surpassing its baseline (Xu et al., 2023) across indoor and outdoor scenes. Applying our framework to the stronger foundation-prior backbones MonSter (Cheng et al., 2025) and FoundationStereo (Wen et al., 2025) further pushes performance to the state-of-the-art, with MAGIC-FoundationStereo ranking first on most metrics. These results demonstrate the effectiveness of our framework across different stereo backbones.

Comparison in occluded regions. As reported in Table 1, MAGIC-IGEV achieves relative outlier reductions of 27.28%, 36.16%, 3.38%, and 18.24% on KITTI 2015, KITTI 2012, Middlebury, and ETH3D, respectively. Qualitative comparisons in Fig. 4 further show that existing supervised methods (Xu et al., 2023; Wang et al., 2024) often produce erroneous disparities in occluded regions, whereas our method recovers more coherent structures. Together, these quantitative and qualitative results show that our framework improves disparity estimation in occluded regions. We additionally analyze the contribution of occlusion-aware weighting in Section 4.4.

Comparison with multi-view methods. NeRFStereo (Tosi et al., 2023) and SMV (Yuan et al., 2021) also exploit multi-view observations for unsupervised training. As shown in Table 1, MAGIC-IGEV consistently outperforms NeRFStereo across all datasets, with particularly pronounced improvements in occluded regions. Since SMV provides no public weights, we follow its evaluation protocol and report zero-shot D1-all on the KITTI 2015 training set, where MAGIC-FoundationStereo achieves 2.57%, compared with 3.82% reported by SMV. MAGIC combines supervision across different baselines with occlusion-aware weighting to improve disparity completion learning in occluded regions.

Cross-weather evaluation. Table 2 reports performance across diverse weather conditions. Rainy scenes are especially challenging, with FoundationStereo's outlier rate rising to 28.50%. In contrast, the diversity of our training data keeps MAGIC-FoundationStereo at 4.18%, a relative reduction of over 85%. Furthermore, although fog is absent from our training data, MAGIC improves every backbone by at least 29%, indicating that our supervision generalizes beyond the training distribution. Qualitative results in Fig. 5 further illustrate that MAGIC yields stable predictions in challenging regions, including road surfaces covered with water.

4.3 Target-Domain Adaptation

After pre-training with geometric consistency, we evaluate unsupervised adaptation on KITTI 2015 and 2012 under both binocular and multi-baseline target-domain fine-tuning settings.

Comparison on KITTI benchmarks. For a fair comparison with existing unsupervised methods, we adopt the standard binocular fine-tuning setting, in which the teacher and student observe identical stereo inputs. As shown in Table 3, MAGIC-IGEV* establishes a new state-of-the-art among unsupervised methods, surpassing the previous best method DualNet (Wang et al., 2025b). We attribute this to the strong generalization of our pre-trained model, which already outperforms several fine-tuned unsupervised methods without any target-domain fine-tuning.

Extension to multi-baseline data. We simulate the multi-baseline setting by synthesizing novel views through warping and inpainting (Wang et al., 2025a). As shown by MAGIC-IGEV§ in Table 3,

Table 4: **Ablation studies** on component design. Row (a) is the SceneFlow-pretrained IGEVStereo baseline. "+" denotes continued training for 200K steps on the indicated data. We report the outlier rate in occluded regions on the KITTI 2015 training set. MBS20K-CN: the clear-noon subset of MBS20K. PTF: photometric threshold filtering. AMF: auto-masking filtering (Godard et al., 2019). OAW: the proposed occlusion-aware weighting strategy for balancing supervision.

	Training Set	Supervision Signal	OCC
(a)	SceneFlow	ground-truth	14.26
(b)	+ MBS20K	ground-truth	<u>10.75</u>
(c)	+ MBS20K	photometric consistency	97.44
(d)	+ MBS20K	binocular self-training	15.55
(e)	+ MBS20K	MAGIC	**10.37**
(f)	+ MBS20K-CN	MAGIC	11.25

Training set vs. supervision signal

	$\mathcal{L}_g$	$\mathcal{L}_p$	$\mathcal{L}_s$	EMA	OCC
(g)		✓	✓		97.44
(h)	✓	✓		✓	<u>10.62</u>
(i)	✓		✓	✓	11.23
(j)	✓	✓	✓		10.87
(k)	✓	✓	✓	✓	**10.37**

Loss terms and EMA teacher

	PTF	AMF	OAW	OCC
(l)				13.67
(m)	✓			13.49
(n)		✓		13.27
(o)	✓	✓		<u>12.39</u>
(p)	✓	✓	✓	**10.37**

Filtering and weighting

this setting achieves the lowest outlier rate, improving over the standard binocular setting. These results show that MAGIC benefits from additional target views during target-domain adaptation.

Comparison on KITTI training sets. Following prior works, we evaluate our models fine-tuned with multi-baseline data (MAGIC-IGEV[§]) on the KITTI training sets (Fig. 6), outperforming TS-PSMNet (Yue et al., 2024) and CST-Stereo (Zhou et al., 2025).

4.4 Ablation Study

Training set vs. supervision signal. In Table 4 (left), continued baseline training with ground-truth supervision on MBS20K (a vs. b) markedly improves generalization, confirming the value of our dataset. On the same MBS20K training set, both photometric supervision (c) and standard binocular self-training (d) degrade relative to the baseline (a). In contrast, MAGIC (e) leverages visibility asymmetry to derive reliable occlusion supervision, achieving a lower occluded-region outlier rate than the ground-truth-supervised baseline (b). Training on the clear-noon subset alone also improves occluded-region disparity estimation over the baseline (f vs. a), indicating that the gains cannot be attributed solely to the weather and illumination diversity of MBS20K.

Loss terms and EMA teacher. In the middle panel, removing the geometric consistency loss $\mathcal{L}_g$ (g) severely degrades occlusion performance because the remaining appearance-based terms provide no direct occlusion supervision. With $\mathcal{L}_g$ retained, incorporating only the photometric term (h) or only the smoothness term (i) remains suboptimal, while jointly optimizing all three terms (k) achieves the best performance. Comparing (j) and (k), the EMA teacher further improves performance by providing more stable and progressively refined pseudo-labels.

Filtering and weighting. In the right panel, starting from the base variant (l), PTF (m) and AMF (n) improve supervision quality by filtering out uncertain pixels, while combining both (o) yields additional improvements. Adding OAW (o vs. p) gives the largest gain in this ablation (12.39% to 10.37%) by strengthening sparse student-occluded yet teacher-visible supervision.

5 Conclusion

We presented MAGIC, a multi-baseline geometric consistency framework for learning disparity completion in occluded regions. Visibility asymmetry enables the teacher to observe correspondences unavailable to the student, while baseline scaling aligns the teacher's disparity predictions to supervise the student's training. Occlusion-aware weighting discards unreliable teacher guidance and strengthens sparse student-occluded yet teacher-visible supervision during learning. We also introduced MBS20K, a synthetic multi-baseline dataset spanning diverse scenes, weather, and illumination. Pre-training on MBS20K improves occlusion performance across backbones and generalization to real-world scenes and weather. On KITTI, the pre-trained model already outperforms several fine-tuned unsupervised methods. Further fine-tuning on binocular pairs achieves state-of-the-art unsupervised performance, with additional gains when more target views are available. Future work will extend training to real multi-camera observations, scaling stereo learning without ground-truth disparity labels across diverse real-world scenes and challenging imaging conditions.

AI USE STATEMENT

We used LLM-based assistants for literature retrieval and discovery, translation, and language polishing of author-written text. We also used these tools to write and edit plotting code and assist with LaTeX typesetting and figure and table layout. The methodology and experimental design were developed by the authors. LLM-based assistants were not used to generate manuscript drafts or synthetic datasets, implement the proposed method, or interpret experimental results. Assistance with mathematical proofs is not applicable to this work. The authors reviewed all AI-assisted work and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The method and training objectives are described in Section 3, and the implementation details and evaluation protocols are provided in Section 4.1. Additional experimental and dataset details are included in the appendix. We provide representative samples from MBS20K as supplementary material; the full dataset is too large to include in the submission package. The source code will be publicly released upon acceptance.

REFERENCES

Filippo Aleotti, Fabio Tosi, Li Zhang, Matteo Poggi, and Stefano Mattoccia. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In *European Conference on Computer Vision*, pp. 614–632. Springer, 2020.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5410–5418, 2018.

Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6273–6282, 2025.

Yikang Ding, Qingtian Zhu, Xiangyue Liu, Wentao Yuan, Haotian Zhang, and Chi Zhang. KD-MVS: Knowledge distillation based self-supervised learning for multi-view stereo. In *Computer Vision – ECCV 2022*, pp. 630–646. Springer, 2022.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3273–3282, 2019.

Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21857–21867, 2025.

Zijie Jiang and Masatoshi Okutomi. Emr-msf: Self-supervised recurrent monocular scene flow exploiting ego-motion rigidity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 69–78, 2023.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pp. 66–75, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL `https://arxiv.org/abs/1412.6980`.

Ang Li and Zejian Yuan. Occlusion aware stereo matching via cooperative unsupervised learning. In *Asian Conference on Computer Vision*, pp. 197–213. Springer, 2018.

Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16263–16272, 2022.

Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pp. 218–227. IEEE, 2021.

Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6648–6657, 2020.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4040–4048, 2016.

Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3061–3070, 2015.

Jiahao Pang, Wenxiu Sun, Chengxi Yang, Jimmy Ren, Ruichao Xiao, Jin Zeng, and Liang Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2070–2079, 2018.

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pp. 31–42. Springer, 2014.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3260–3269, 2017.

Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 231–240, 2017.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13906–13915, 2021.

Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018. URL https://arxiv.org/abs/1803.09820.

Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: a simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10328–10337, 2021.

Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 195–204, 2019.

Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 855–866, 2023.

Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. Weakly supervised learning of deep metrics for stereo reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1339–1348, 2017.

Haiyang Wang, Xinchao Wang, Jie Song, Jie Lei, and Mingli Song. Faster self-adaptive deep stereo. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2108–2125, 2022.

Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19701–19710, 2024.

Xianqi Wang, Hao Yang, Gangwei Xu, Junda Cheng, Min Lin, Yong Deng, Jinliang Zang, Yurui Chen, and Xin Yang. Zerostereo: Zero-shot stereo matching from single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28177–28187, 2025a.

Yun Wang, Jiahao Zheng, Chenghao Zhang, Zhanjie Zhang, Kunhong Li, Yongjian Zhang, and Junjie Hu. Dualnet: Robust self-supervised stereo matching with pseudo-label supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8178–8186, 2025b.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5249–5260, 2025.

Bowen Wen, Shaurya Dewan, and Stan Birchfield. Fast-foundationstereo: Real-time zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7513–7524, 2026.

Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21919–21928, 2023.

Peng Xu, Zhiyu Xiang, Chengyu Qiao, Jingyun Fu, and Tianyu Pu. Adaptive multi-modal cross-entropy loss for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5135–5144, 2024.

Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 899–908, 2019.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

Ruizhi Yang, Xingqiang Li, Rigang Cong, and Jinsong Du. Unsupervised hierarchical iterative tile refinement network with 3d planar segmentation loss. *IEEE Robotics and Automation Letters*, 9 (3):2678–2685, 2024c.

Weihao Yuan, Yazhan Zhang, Bingkun Wu, Siyu Zhu, Ping Tan, Michael Yu Wang, and Qifeng Chen. Stereo matching by self-supervision of multiscopic vision. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5702–5709, 2021.

Xin Yue, Zongqing Lu, Xiangru Lin, Wenjia Ren, Zhijing Shao, Haonan Hu, Yu Zhang, and Qingmin Liao. Semi-stereo: A universal stereo matching framework for imperfect data via semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 646–655, 2024.

Jiawei Zhang, Jiahe Li, Lei Huang, Xiaohan Yu, Lin Gu, Jin Zheng, and Xiao Bai. Robust synthetic-to-real transfer for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20247–20257, 2024.

Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1327–1336, 2023.

Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.

Jingyi Zhou, Peng Ye, Haoyu Zhang, Jiakang Yuan, Rao Qiang, Liu YangChenXu, Wu Cailin, Feng Xu, and Tao Chen. Consistency-aware self-training for iterative-based stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16641–16650, 2025.

APPENDICES

This document provides additional details and analyses supporting the main paper. Section A reports our computing infrastructure together with the implementation details omitted from the main paper. Section B details our MBS20K dataset, and Section C introduces the construction of the multi-baseline fine-tuning data. We then report quantitative results on the KITTI benchmarks in Section D. Section E quantifies the reliability of our cross-baseline pseudo-labels and the quality of the computed visibility masks. Section F analyzes the sensitivity of our hyperparameters and the robustness of our framework to calibration error. Finally, Section G outlines our plan to extend MAGIC to multi-baseline observations captured with real camera arrays.

A IMPLEMENTATION DETAILS

A.1 COMPUTING INFRASTRUCTURE

All experiments are conducted on a single workstation equipped with four NVIDIA RTX 4090 GPUs (48 GB of memory each), two Intel Xeon Gold 6148 CPUs at 2.40 GHz, and 256 GB of system memory, running Ubuntu 22.04. Our implementation uses Python 3.10 and PyTorch 2.4.1.

A.2 AUTO-MASKING CRITERION

Besides photometric thresholding, the visibility masks adopt the auto-masking criterion of (Godard et al., 2019). A pixel passes this criterion only when warping the target view lowers its photometric error relative to comparing the two images directly,

$$\mathcal{L}_p(I_{ref}, \hat{I}_{ref}) < \mathcal{L}_p(I_{ref}, I_{tgt}). \tag{6}$$

The pixels this criterion removes differ between monocular depth estimation and stereo matching. In the monocular setting the two images are consecutive video frames, and the criterion discards pixels whose appearance does not change with ego-motion, namely pixels in frames captured while the camera is static, and pixels on objects that move at the same velocity as the camera. In stereo matching the two images are captured simultaneously by two cameras with a fixed baseline, so neither case arises. A pixel instead fails Eq. (6) when its disparity is so small that warping leaves it essentially unchanged, or when its neighborhood is so uniform that the warped and unwarped errors are indistinguishable. The former corresponds to points at very large depth, most notably sky. The latter corresponds to textureless surfaces such as blank walls. The supervision in both groups of pixels is unreliable, so they are excluded from supervision.

A.3 ABLATION VARIANTS

We detail the settings of variants (c) and (d) in Table 4 of the main paper, both following the same training protocol used for the full MAGIC framework in row (e).

Row (c). This variant continues training the SceneFlow-pretrained baseline (Xu et al., 2023) on MBS20K with the photometric loss and the edge-aware smoothness term (Zhong et al., 2017) alone, dropping the geometric consistency loss and the teacher branch. It is the same variant as row (g). The photometric term is masked by $\mathcal{V}^s$ to discard erroneous supervision in occluded regions, leaving direct correspondence supervision only in non-occluded pixels. We attribute this variant's degradation to catastrophic forgetting of the completion ability acquired during supervised pre-training.

Row (d). This variant feeds the teacher and the student the same target view, so the baseline ratio becomes one and the geometric consistency loss reduces to conventional binocular self-training. We also set $\omega = 1$, since no student-occluded yet teacher-visible region arises. The occlusion-aware weight map accordingly reduces to the teacher visibility mask $\mathcal{V}^t$, which still discards pseudo-labels where the teacher is occluded. All other settings follow MAGIC, including the EMA teacher, the student-only data augmentation, and the visibility filtering strategy.

Table 5: **Statistics of MBS20K.** The dataset covers diverse weather and illumination conditions. Each set contains five color images and one ground-truth depth map.

Town	Clear Night	Clear Noon	Clear Sunset	Cloudy Noon	Rainy Noon	Total
01	630	630	629	632	631	3152
02	424	424	424	425	424	2121
03	647	649	648	649	648	3241
05	563	563	567	567	564	2824
06	820	821	820	821	821	4103
07	442	440	441	440	440	2203
10	552	557	554	554	556	2773
Total	**4078**	**4084**	**4083**	**4088**	**4084**	**20417**

B MBS20K DATASET

Data collection setup. We collect MBS20K in the CARLA simulator (Dosovitskiy et al., 2017) using a Tesla Model 3 equipped with five synchronized RGB cameras. Data collection spans seven towns (01, 02, 03, 05, 06, 07, and 10). For each town, we first record driving trajectories at clear noon, then replay them under four additional conditions, namely clear sunset, clear night, cloudy noon, and rainy noon. The ego-vehicle operates at speeds up to 30 km/h, with frames captured at 0.5-second intervals. All five cameras are rigidly mounted at a height of 2.26 m, maintaining a uniform baseline of 0.5 m between adjacent sensors to form a fixed multi-baseline array.

Ground-truth depth acquisition. Although our training objective requires no ground-truth disparities, we record metric depth to support broader research uses. A depth camera is co-located with the leftmost RGB camera, storing pixel-aligned depth measurements alongside the RGB images. This inclusion supports fully supervised stereo training, monocular depth estimation, multi-task learning, and quantitative evaluations that require metric depth. Following the standard CARLA configuration (Dosovitskiy et al., 2017), depth maps are encoded as three-channel RGB images. The metric depth z, measured in meters, is decoded from the three color channels as follows:

$$z = 1000 \times \frac{R + 256G + 256^2 B}{256^3 - 1}. \tag{7}$$

Dataset statistics. Table 5 summarizes the distribution of multi-baseline image sets collected across different towns and environmental conditions. In total, MBS20K comprises 20,417 synchronized multi-baseline sets. Representative multi-baseline sets are provided in the accompanying data supplement, showing the diversity in urban layouts, lighting, and weather conditions.

Input triplet sampling strategy. To maximize training diversity, we construct triplets from the N synchronized views available at each timestamp using the following protocol: (1) one image is uniformly selected as the *reference view*; (2) two target views are independently sampled (with replacement) from the remaining $(N-1)$ images to serve as the student and teacher inputs. This strategy generates $N(N-1)^2$ possible triplets per timestamp. With our five-camera setup ($N = 5$), this results in $5 \times 4^2 = 80$ unique triplets per frame. Scaling this across the entire dataset yields a total of $20{,}417 \times 80 = 1{,}633{,}360$ training triplets.

Handling targets on either side of the reference camera. Our sampling strategy allows the target camera to lie on either side of the reference camera, whereas the stereo network assumes a left reference image and a right target image. When the target camera lies to the left, we horizontally flip both images and feed them to the network in reference–target order. This reverses the horizontal correspondence direction, making the transformed pair compatible with the network's convention while preserving the identity of the reference view. The predicted disparity map is then horizontally flipped back to align with the original reference image. We represent disparity as a nonnegative magnitude and use positive baseline lengths; therefore, this final flip changes only the spatial indexing of the disparity map, without changing the sign or magnitude of any disparity value.

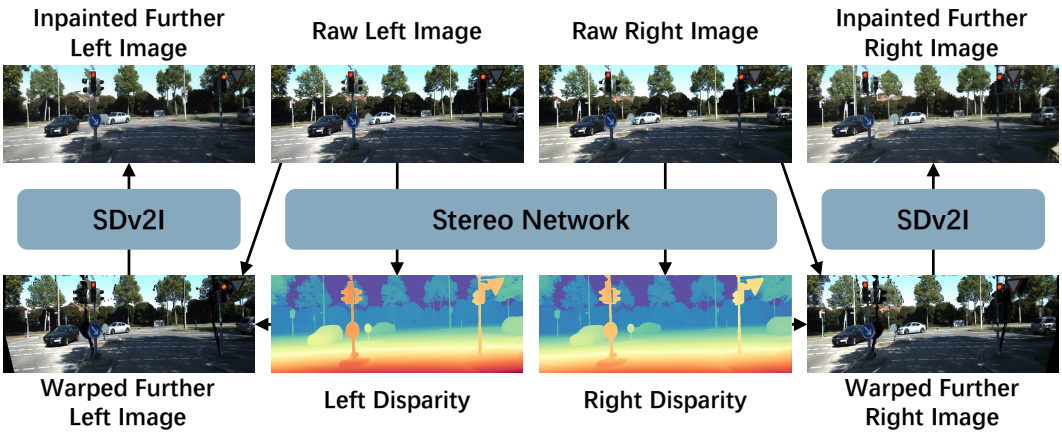


Figure 7: **Pipeline for constructing multi-baseline data from stereo image pairs.** SDv2I: Stable Diffusion v2 Inpainting model (Wang et al., 2025a).

C Construction of Multi-Baseline Fine-Tuning Data

Because KITTI 2012 (Geiger et al., 2012) and KITTI 2015 (Menze & Geiger, 2015) provide only binocular pairs, we synthesize additional views to evaluate multi-baseline fine-tuning in the target domain. Following ZeroStereo (Wang et al., 2025a), we synthesize the additional views by warping and inpainting, but estimate disparity from each real stereo pair rather than from a monocular depth prior (Yang et al., 2024b). The synthesized views are scale-consistent with the true camera baseline.

As illustrated in Fig. 7, the view generation process consists of three steps:

1. **Disparity estimation.** We use MonSter (Cheng et al., 2025) to estimate disparity maps for both images in each original left-right stereo pair.
2. **Geometric warping.** Using the estimated disparity, we generate two additional outer views by warping the left image further to the left and the right image further to the right. The baseline between adjacent views matches the original stereo baseline since all warping operations are performed in the same original disparity coordinate space.
3. **Generative inpainting.** Warping leaves disoccluded holes because some pixels in the warped view have no valid correspondence in the original view. Disparity errors near depth discontinuities can also cause boundary bleeding and distortion (Xu et al., 2024). We therefore dilate the occlusion mask to cover both kinds of artifacts and fill the resulting masked regions with an inpainting model (Wang et al., 2025a).

The output contains four views per timestamp: the stereo pair and two generated outer views. During fine-tuning, triplets sampled from these views provide supervision based on geometric consistency.

Discussion. MAGIC-IGEV[§] relies on generated views as a proxy for real multi-baseline observations. Warping and inpainting artifacts in these generated views add noise to the supervision. The reported gains concern this proxy setting. Real camera arrays introduce different acquisition errors, so the benefits of physically captured views cannot be inferred directly from these proxy results.

D Additional KITTI Benchmark Results

As shown in Table 6, under binocular fine-tuning, MAGIC-IGEV* surpasses all previous unsupervised methods on most reported metrics. Binocular fine-tuning gives the teacher and the student the same target view, so we set $\omega = 1$ as in row (d) and the occlusion-aware weight map reduces to the teacher visibility mask $\mathcal{V}^t$. The official leaderboard does not report occluded-region metrics. Compared with DualNet (Wang et al., 2025b), however, our Out-3 on KITTI 2012 (Geiger et al., 2012) decreases only marginally on NOC pixels, from 2.06% to 2.02%, but far more on ALL pixels, from 2.59% to 2.36%, which indicates that the gain concentrates on occluded pixels. We attribute

Table 6: **Additional results on the KITTI benchmarks**. This table extends Table 3 of the main paper with more evaluation metrics and baseline methods. The first block lists supervised methods, separated by a double rule from the unsupervised methods below. In the last block, the unmarked MAGIC row denotes our pre-trained model without fine-tuning, * binocular fine-tuning, and § multi-baseline fine-tuning, following the main paper. # follows the settings of § but trains from random initialization instead of the official SceneFlow-pretrained weights. **Bold** and underline mark the best and second-best results among unsupervised methods only.

Method	KITTI 2015						KITTI 2012					
	NOC			ALL			NOC			ALL		
	D1-BG	D1-FG	D1-ALL	D1-BG	D1-FG	D1-ALL	EPE	Out-2	Out-3	EPE	Out-2	Out-3
DispNet (2016)	4.11	3.72	4.05	4.32	4.41	4.34	0.9	7.38	4.11	1.0	8.11	4.65
SGM-Net (2017)	2.23	7.44	3.09	2.66	8.64	3.66	0.7	3.60	2.29	0.9	5.15	3.50
GCNet (2017)	2.02	5.58	2.61	2.21	6.16	2.87	0.6	2.71	1.77	0.7	3.46	2.30
MADNet (2019)	3.45	8.41	4.27	3.75	9.20	4.66	—	—	—	—	—	—
AdaStereo (2021)	2.39	5.06	2.83	2.59	5.55	3.08	—	—	—	—	—	—
DualNet (2025b)	2.28	4.66	2.67	2.46	5.25	2.92	0.6	—	2.06	**0.6**	—	2.59
MC-CNN-WS (2017)	3.06	9.42	4.11	3.78	10.93	4.97	0.8	4.76	3.02	1.0	6.57	4.45
SsSnet (2017)	2.46	6.13	3.06	2.70	6.92	3.40	0.7	3.34	2.30	0.8	4.24	3.00
OASM (2018)	5.44	17.30	7.39	6.89	19.42	8.98	1.3	9.01	6.39	2.0	11.17	8.60
Flow2Stereo (2020)	4.77	14.03	6.29	5.01	14.62	6.61	1.0	6.56	4.58	1.1	7.32	5.11
Reversing-PSM (2020)	2.97	8.33	3.86	3.13	8.70	4.06	—	—	—	—	—	—
PASMnet (2022)	5.02	15.16	6.69	5.41	16.36	7.23	1.3	—	7.14	1.5	—	8.57
EMR-MSF (2023)	8.30	14.16	9.27	8.61	15.15	9.70	—	—	—	—	—	—
UHP (2024c)	4.65	12.37	5.93	5.00	13.70	6.45	1.2	9.08	6.05	1.3	10.37	7.09
MAGIC-IGEV	2.67	12.36	4.27	2.89	13.19	4.60	0.7	4.18	2.72	0.8	5.05	3.27
MAGIC-IGEV*	2.13	4.94	2.59	2.28	5.39	2.79	0.6	3.23	2.02	**0.6**	3.79	2.36
MAGIC-IGEV§	**2.06**	4.43	**2.45**	**2.21**	4.86	**2.65**	**0.5**	**3.11**	**1.92**	**0.6**	**3.69**	**2.31**
MAGIC-IGEV#	2.25	**4.30**	2.59	2.44	**4.75**	2.82	0.6	3.54	2.17	**0.6**	4.21	2.60

Table 7: **Reliability of cross-baseline pseudo-labels.** Region (2) is teacher-occluded and region (3) is student-occluded yet teacher-visible, following the definitions in the method section of the main paper. We compare the rescaled teacher pseudo-label $r \cdot d^t$ against the student prediction d^s, both measured against the ground-truth disparity at the student baseline. Region (2) reports the teacher alone, since these pixels are discarded rather than compared. The upper block uses oracle visibility derived from ground-truth depth; the lower block uses photometric thresholding and auto-masking of fixed-model predictions, without training or weight updates.

Partition	Teacher pseudo-label		Student prediction	
	EPE	Out-3	EPE	Out-3
Oracle visibility				
Region (2)	7.29	32.22	—	—
Region (3)	**4.27**	**19.72**	5.19	24.14
Computed visibility				
Region (2)	7.66	28.23	—	—
Region (3)	**2.27**	**12.91**	3.52	17.75

this to geometrically consistent pre-training across multiple baselines, which strengthens the ability to complete disparities within the occluded image regions.

MAGIC-IGEV§ achieves the best results on most reported metrics, further supporting the use of multi-baseline observations for fine-tuning within the target domain.

Random initialization. To assess dependence on SceneFlow initialization, we also report MAGIC-IGEV# with random initialization. As described in the main paper, photometric supervision anchors basic matching, teacher pseudo-labels cover regions where photometric consistency fails, and the EMA update propagates the student's improved disparity completion back to the teacher. This

Table 8: **Mask quality and region size** across the six camera pairs, all reported in percent. $\mathcal{V}^t$ precision is the fraction of pixels marked as teacher-visible by the computed mask that are genuinely teacher-visible, so its complement measures the fraction of admitted pixels that are actually teacher-occluded. Occlusion recall is the fraction of genuinely teacher-occluded pixels that $\mathcal{V}^t$ discards, so rows sharing the same B^t report identical values. The last column additionally reports how much of the reference image falls in region (3), the pixels occluded to the student yet visible to the teacher.

$B^t \to B^s$ (m)	$\mathcal{V}^t$ precision	Occ. recall	Region (3)
$0.5 \to 1.0$	98.8	87.3	5.8
$0.5 \to 1.5$	98.8	87.3	9.8
$0.5 \to 2.0$	98.8	87.3	13.5
$1.0 \to 1.5$	98.5	91.7	5.5
$1.0 \to 2.0$	98.5	91.7	9.5
$1.5 \to 2.0$	98.4	93.6	5.5

complementary objective enables training from random initialization while remaining competitive with variants initialized from official SceneFlow-pretrained weights (Mayer et al., 2016).

E Analysis of Occlusion-Aware Weighting

Our framework exploits the premise that teacher-visible pixels provide more reliable predictions on average where the student is occluded. We evaluate this premise using the pixel-aligned ground-truth depth maps recorded by the depth camera of our MBS20K dataset.

Setup. We evaluate 500 frames covering all five weather conditions at the full 540×960 resolution. The reference view is fixed to the leftmost camera so that the co-located depth camera supplies ground-truth disparity, and we enumerate the six camera pairs satisfying $B^t < B^s$, the regime in which region (3) arises. Oracle visibility is obtained by forward-warping the ground-truth disparity with a z-buffer test, which marks a pixel occluded when a nearer pixel maps to the same target pixel or when it falls outside the target view, both cases lacking a valid correspondence. A single IGEVStereo model (Xu et al., 2023) with official SceneFlow weights (Mayer et al., 2016) serves as both the teacher and the student, so the two branches differ only in their target view and the resulting baseline ratio, and the diagnostic involves no MAGIC training.

Teacher visibility yields more reliable supervision. As reported in Table 7, the teacher pseudo-label is more accurate than the student prediction on average in region (3), by 0.92 px EPE under the oracle partition and by 1.25 px under the computed partition. The advantage holds even though rescaling by $r > 1$ amplifies the teacher's error proportionally. In region (2) the teacher pseudo-label is far less accurate than in region (3), with Out-3 above 28% under both partitions, which supports discarding it. Both partitions agree on the direction of every comparison, while photometric thresholding and auto-masking yield a computed partition with lower absolute errors.

Mask quality and region size. As reported in Table 8, the computed $\mathcal{V}^t$ has 98.4–98.8% precision: only 1.2–1.6% of the pixels admitted into $\mathcal{L}_g$ are actually teacher-occluded. At the same time, $\mathcal{V}^t$ filters out 87.3–93.6% of the genuinely teacher-occluded pixels. Region (3) covers 5.5% to 13.5% of the reference image, a small fraction compared with the co-visible region. The supervision unique to visibility asymmetry is therefore sparse, motivating the weight $\omega > 1$.

F Sensitivity and Robustness

F.1 Hyperparameter Search

We conduct a grid search over λ_p, τ, and ω, training on MBS20K and evaluating on the KITTI 2015 training set (Menze & Geiger, 2015). As shown in Fig. 8, incorporating photometric supervision improves performance. However, excessively large λ_p introduces noise due to imperfect occlusion masking. Similarly, a loose threshold τ fails to adequately filter occluded pixels and leads to degraded performance. For the occlusion-aware weighting factor ω, a moderate value promotes disparity completion in student-occluded but teacher-visible regions, whereas an excessive value amplifies the noise in the teacher pseudo-labels. We select $\lambda_p = 10$, $\tau = 0.1$, and $\omega = 2$ using

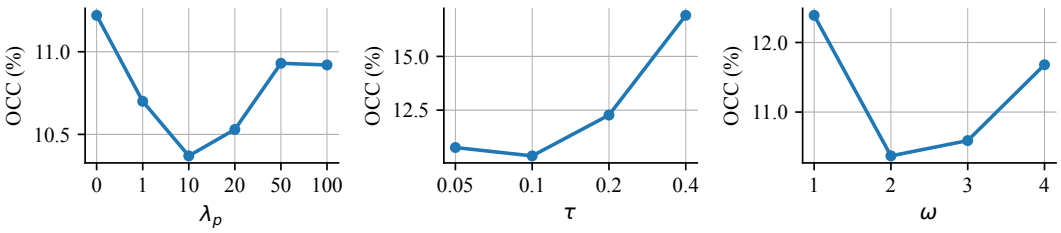


Figure 8: **Hyperparameter search.** We report Out-3 error in occluded regions on the KITTI 2015 training set (Menze & Geiger, 2015) while varying λ_p, τ, and ω.

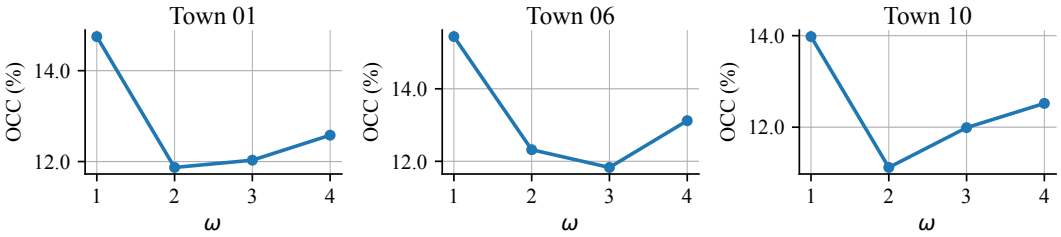


Figure 9: **Sensitivity of ω across scene layouts.** Each panel trains on a single town of MBS20K and reports Out-3 error in occluded regions on the KITTI 2015 training set (Menze & Geiger, 2015) for different values of the occlusion-aware weighting parameter ω.

KITTI 2015 training-set annotations for hyperparameter selection only. We then evaluate cross-dataset generalization on KITTI 2012 (Geiger et al., 2012), Middlebury (Scharstein et al., 2014), ETH3D (Schops et al., 2017), and DrivingStereo (Yang et al., 2019) with these hyperparameters fixed, without using annotations from these four datasets for training or hyperparameter selection.

F.2 Sensitivity of ω Across Scenes

The proportion of occluded pixels varies substantially across scenes, so a practical ω should generalize across them without retuning. The search above trains on all towns of MBS20K at once, so its ω reflects an average over layouts. We therefore train on three towns with markedly different layouts, a small town (town 01), a rural road (town 06), and a high-rise city center (town 10), whose occlusion structures differ accordingly, and sweep ω alone. As shown in Fig. 9, every value of ω between 2 and 4 outperforms $\omega = 1$ in all three towns, and $\omega = 2$ ranks best or second best in each. These results support emphasizing the student-occluded yet teacher-visible region and using $\omega = 2$ across different training layouts without retuning for each town.

F.3 Transfer of the Photometric Threshold

The threshold $\tau = 0.1$ is used throughout pre-training on the synthetic MBS20K, whose appearance differs substantially from real imagery. Since the photometric error depends on appearance, we examine whether τ still separates occluded from non-occluded pixels on real data. Middlebury (Scharstein et al., 2014) and ETH3D (Schops et al., 2017) provide ground-truth disparity and occlusion masks, which lets us warp the right image to the left view with ground-truth disparity and histogram the photometric error separately over the two populations. As shown in Fig. 10, the two populations remain well separated on both datasets, with the non-occluded pixels concentrated near zero and the occluded pixels spread over much larger errors. At $\tau = 0.1$ the threshold retains 80.2% and 86.1% of the non-occluded pixels while filtering 92.1% and 84.0% of the occluded pixels on Middlebury and ETH3D, respectively, so the value used in synthetic pre-training transfers to both real indoor and outdoor imagery without retuning the photometric threshold.

F.4 Robustness to Calibration Error

Real multi-camera arrays carry extrinsic calibration error. To verify how much baseline error our framework tolerates, we perturb the ratio $r = B^s/B^t$, the only quantity in the geometric consistency

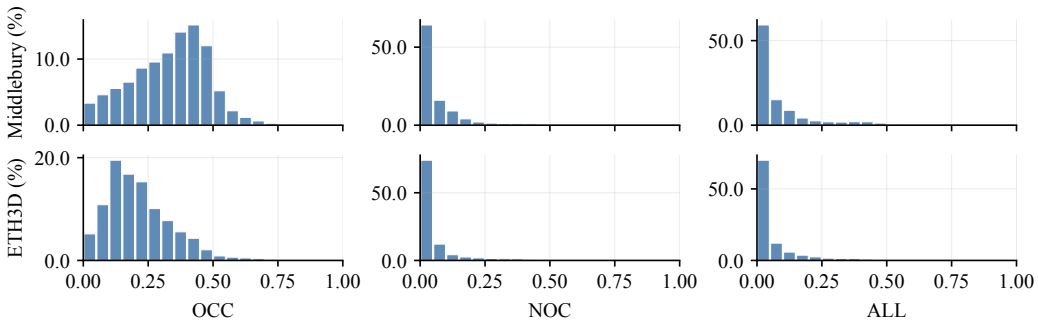


Figure 10: **Distribution of photometric error on real-world data.** We warp the right image to the left view with ground-truth disparity on Middlebury (Scharstein et al., 2014) and ETH3D (Schops et al., 2017) and histogram the resulting photometric error over occluded (OCC), non-occluded (NOC), and all (ALL) pixels. The horizontal axis is the photometric error and the vertical axis represents the percentage of pixels within each histogram bin.

Table 9: **Robustness to baseline-ratio perturbation.** We perturb the baseline ratio during training as $r' = r(1+\epsilon)$ with $\epsilon \sim \mathcal{U}(-\alpha, \alpha)$ resampled for every training sample, and report Out-3 error on the KITTI 2015 training set in occluded (OCC), non-occluded (NOC), and all (ALL) regions.

α	0	0.5%	1%	2%	5%
OCC	10.37	10.27	10.46	10.53	11.14
NOC	4.05	4.02	4.07	4.10	4.25
ALL	4.21	4.18	4.23	4.27	4.42

loss that depends on calibration. During training, we sample $r' = r(1+\epsilon)$, where $\epsilon \sim \mathcal{U}(-\alpha, \alpha)$. As reported in Table 9, Out-3 in all three regions stays within 0.16 percentage points of the unperturbed setting for α up to 2%. At $\alpha = 5\%$ the degradation remains graceful, with the occluded-region error rising from 10.37% to 11.14%. The photometric term may contribute to this stability, since it anchors the student to appearance-based matching independently of the baseline ratio r.

G Future Work

All multi-baseline observations used in this work are either rendered in simulation or synthesized from binocular pairs. The perturbation study of Section F.4 probes the sensitivity of MAGIC to inaccurate baseline measurements, but it cannot replace physically captured data. We therefore plan to build a real multi-baseline capture rig and collect a dataset spanning diverse scenes, illumination, and weather conditions. Such data would let us study the acquisition artifacts that simulation does not reproduce, including inter-camera exposure and color mismatch as well as imperfect synchronization under ego-motion. Our framework requires only prior calibration of the array baselines, and needs neither an expensive LiDAR sensor nor any other modality. We aim to use low-cost camera arrays to close the synthetic-to-real gap and scale unsupervised stereo pre-training to real data.