/TemplateVersion (2026.1)

利用多基线对比学习进行自监督立体匹配

Peng Xu, Zhiyu Xiang*, Jingyun Fu, Tianyu Pu, Kai Wang,
Chaojie Ji, Tingming Bai, Eryun Liu

College of Information Science and Electronic Engineering, Zhejiang University, China
xxxupeng@zju.edu.cn

Abstract

当前的自监督立体匹配依赖于光度一致性假设，这在由于对应关系不明确而导致的遮挡区域中失效。为了解决这个问题，我们提出了 BaCon-Stereo，这是一种简单而有效的对比学习框架，用于非遮挡和遮挡区域中的自监督立体网络训练。我们采用一种教师-学生范式，用多基线输入，其中传递给教师和学生的立体图像对共享相同的参考视图，但在目标视图上不同。从几何上讲，学生目标视图中被遮挡的区域通常在教师的视图中可见，这使得教师在这些区域的预测变得更容易。教师的预测被缩放以匹配学生的基线，然后用于监督学生。我们还引入了一个遮挡感知的注意力图，以更好地指导学生学习遮挡填充。为支持训练，我们合成了一个多基线数据集 BaCon-20k。大量实验表明，BaCon-Stereo 在遮挡和非遮挡区域中均提高了预测性能，具备强大的泛化能力和鲁棒性，并且在 KITTI 2015 和 2012 基准测试上优于最先进的自监督方法。我们的代码和数据集将在论文接受后发布。

作为计算机视觉中的经典主题之一，立体匹配旨在通过估计参考图像和目标图像中对应像素之间的水平位移（i.e.，视差）来恢复三维几何。这在自动驾驶、增强现实和机器人领域中起着至关重要的作用。近年来，全监督的立体匹配取得了显著的进展。然而，在真实世界数据上训练立体网络仍然具有挑战性，因为这通常需要代价高昂的基于 LiDAR 的视差注释。

在这种背景下，自监督立体匹配成为一个有吸引力的解决方案，特别是在特定真实场景中进行微调。自监督方法通常依赖于光度一致性作为监督信号 (Zhong, Dai, and Li 2017)。然而，参考图像和目标图像之间的非重叠视野，以及前景遮挡，通常导致某些参考像素在目标图像中不存在。这些像素违反了光度一致性假设，使得网络在遮挡区域内天真地复制邻近视差，以最小化光度损失，如 Fig. 1 (a) 所示。

最近的一些研究尝试检测遮挡的像素，并将其从光度损失中排除，从而防止网络在训练过程中被误导。例如，OASM (Li and Yuan 2018) 直接使用一个子网络从参考图像中预测遮挡掩码。PASMnet (Wang et al. 2022) 在其中间注意力图上进行阈值处理，以去除所有候选视差中具有低特征相似性的像素。尽管这些方法提高了匹配性能， Fig. 1 (b) 显示，仅在非遮挡区域进行训练不足以在遮挡区域产生可靠的视差预测。我们认为，遮挡区域的准确监督对于使立体网络学习有效的视差填充行为至关重要。

*Corresponding author.

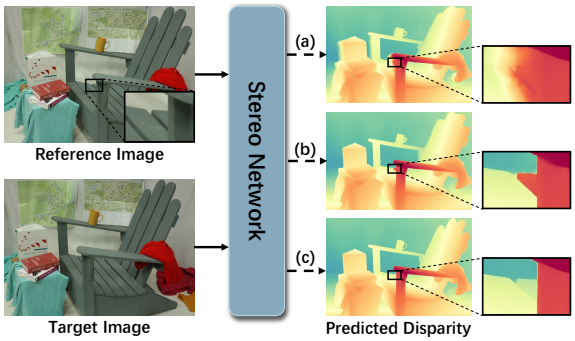

Figure 1: 遮挡区域中视差预测的示例。(a) 仅通过光度损失训练的立体网络在没有遮挡掩码的情况下，利用“捷径”通过从相邻未遮挡区域填充视差来处理遮挡区域。(b) 使用遮罩的光度损失训练的立体网络在遮挡区域做出不正确的预测。(c) 使用我们提出的遮挡感知对比损失训练的立体网络在未遮挡和遮挡区域分别实现了准确的匹配和补全。

在这项工作中，我们利用额外的目标线索来获得遮挡区域的可靠伪真值。具体而言，我们通过复制训练过的学生网络来实例化一个教师网络。教师和学生网络共享相同的参考图像，但输入不同的目标图像。由于目标视角和基线长度的变化，它们的遮挡区域和视差尺度不同。然而，它们仍需预测一致的参考几何。我们将教师输出的视差重新缩放以匹配学生的基线，并将其用作伪真值来监督学生。同时，我们屏蔽掉那些教师遮挡的像素，因这些区域的伪真值具有高度不确定性。监督信号覆盖了共享的非遮挡区域，以及教师友好但学生挑战的区域。通过这种多基线配置，我们在遮挡和非遮挡区域对学生进行训练监督。

此外，我们观察到，在整个图像上施加统一的惩罚未能充分地监督遮挡区域中的立体网络。神经网络倾向于首先学习简单和清晰的模式 (Arpit et al. 2017)，这意味着立体网络偏向于学习简单的匹配而不是具有挑战性的填补。我们通过遮挡感知监督来缓解这个问题，该监督在对学生不可见但对教师可见的区域增加损失，驱动学生主动发展遮挡填补能力。最后，通过反向传播更新学生，而教师的参数通过学生参数的指数移动平均来更新。我们从 CARLA 模拟器 (Dosovitskiy et al. 2017) 中合成了多基线立体数据集 BaCon-20k 用于训练，并从真实世界的双目图像生成新视图用于微调。大量的消

融研究验证了我们方法的有效性。此外，我们的方法在零样本泛化和微调性能上效果显著。我们的主要贡献如下：

- 我们提出了一种新颖的多基线对比学习框架，结合动量教师，使得基于几何一致性在遮挡和非遮挡区域都能够进行自监督立体网络训练。
- 我们设计了一个遮挡感知注意力图，以引导立体网络在遮挡区域学习强大的补全能力。
- 我们合成了一个涵盖多种光照和天气条件的多基线立体数据集，以支持多基线对比学习。
- 我们的方法在 KITTI 2015 和 2012 基准测试中的自监督立体方法中达到新的最先进的性能。
- 我们的方法显著提升了现有立体网络的零样本泛化能力，并增强了在不同天气条件下的鲁棒性。

相关工作

深度立体匹配。在过去的几十年中，立体匹配经历了快速的发展，从手工设计的算法（Scharstein and Szeliski 2002）转变为半学习方法（Žbontar and LeCun 2016），最后发展为端到端学习方法（Mayer et al. 2016）。当前的深度立体方法主要分为以下几类：基于 3D 卷积的代价聚合（Chang and Chen 2018; Chen et al. 2022; Xu et al. 2025）和基于 ConvGRU 的迭代优化（Lipson, Teed, and Deng 2021; Li et al. 2022）。IGEVStereo（Xu et al. 2023）采用轻量级的 3D 卷积网络，为迭代单元提供高质量的初始视差。Selective-Stereo（Wang et al. 2024）通过跨多个频率聚合隐藏信息，以提高边缘和平滑区域的精度。DEFOM-Stereo（Jiang et al. 2025）利用单目深度基础模型（Yang et al. 2024b）的强大表示，取得了令人印象深刻的性能。

自监督立体匹配。自监督立体匹配作为一种有前途的范例出现，以克服对昂贵视差标注的依赖。SsSnet（Zhong, Dai, and Li 2017）通过结合光度一致性和视差平滑性来奠定基础，以在未标记的立体对上进行训练。在此基础上，OASM（Li and Yuan 2018）引入了一个遮挡推断模块，以过滤掉不符合光度一致性假设的区域。PVStereo（Wang et al. 2021）利用金字塔投票方案聚合多尺度视差假设，生成可靠的伪地面真值进行训练。PASMnet（Wang et al. 2022）对其中间注意力图应用阈值，以过滤掉通常出现在遮挡区域中的低置信度匹配。UHP（Yang et al. 2024c）使用 3D 平面线索减轻光度损失的错误指导问题。

对比学习。对比学习旨在通过在嵌入空间中将正样本拉近同时推开负样本来学习表示。InstDisc（Wu et al. 2018）在该方向上开创了先河，引入了一个外部记忆库来存储和采样负样本。MoCo（He et al. 2020）用一个动态队列替换了记忆库，并引入了通过指数移动平均更新的动量编码器，这首次在下游任务上超越了完全监督的对应方法。BYOL（Grill et al. 2020）提出了一个无需负样本的新范式，并使用非对称网络来防止模型崩溃。DINO（Caron et al. 2021）用视觉变换器替代卷积骨干网络，并对教师模型应用中心化归一化以增强训练稳定性。

DualNet（Wang et al. 2025）引入了无负对比学习到立体匹配中，该方法通过在第一阶段中通过特征度量一致性训练教师模型，然后在第二阶段中使用其监督学生的概率分布。然而，在 DualNet 中，教师模型仍然无法在遮挡区域提供准确的监督。相反，我们的方法将不同的目标视图输入到学生模型和动量教师模型中，同时加强像素级几何一致性，从而在单次训练阶段内有效地监督学生模型中的遮挡部分。

方法

预备知识

给定一对校正过的立体图像，立体匹配的目标是为参考图像 I_{ref} 中的每个像素在目标图像 I_{tgt} 中找到对应的像素。对应像素之间的水平位移称为视差 d。通过三角测量，可以计算出参考图像中每个像素的深度 z：

$$z = \frac{B \cdot f}{d} \tag{1}$$

，其中 B 是基线长度，i.e. 是两个相机光学中心之间的距离，f 是焦距。

光度一致性假设通常用于自监督立体网络的训练中（Zhong, Dai, and Li 2017）。目标图像 I_{tgt} 通过基于视差的变形被重建为参考图像 $\hat{I}_{ref}$。然后计算 I_{ref} 和 $\hat{I}_{ref}$ 之间的光度差距，作为 SSIM 项（Wang et al. 2004）和 L_1 范数项的加权和：

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

边缘感知平滑损失（Zhong, Dai, and Li 2017）有助于在非边缘区域中鼓励平滑视差，同时在边缘处保持锐利的过渡：

但是，光度一致性假设具有内在的局限性：它依赖于一个前提，即跨视图的对应像素具有相似的外观。这个前提在被遮挡的区域中常常失效，因为在这些区域不存在有效的对应关系，从而误导了监督信号。此外，仅凭平滑损失不足以在这些具有挑战性的区域提供充分的监督。为了解决这一限制，我们利用不同基线捕获的立体图像对之间的几何一致性，甚至在光度提示失效的区域也能提供有效的监督。

多基线对比学习框架

在本节中，我们介绍我们的多基线对比学习框架。为清晰起见，在下面的描述中，我们使用上标 s 和 t 分别表示学生和教师网络。

如图 Fig. 2 所示，我们将三视图图像输入到教师和学生网络中以预测视差图。尽管教师和学生的输入对可能具有不同的基线长度，导致预测的视差图在不同尺度上，但共享的参考图像捕捉到相同的绝对深度。根据 Eq. (1)，我们有如下等式：

$$z = \frac{B^s \cdot f}{d^s} = \frac{B^t \cdot f}{d^t} \tag{3}$$

学生网络。我们仅对学生的输入图像进行数据增强，包括颜色抖动和随机遮挡。通过设置更具挑战性的优化目标（Yang et al. 2024a），学生网络被迫学习更强大和更稳健的特征表示，以保持准确的视差预测。

教师网络。教师网络用与学生网络相同的权重初始化。在训练期间，教师的梯度被截断，其权重 θ^t 在每次迭代中由学生的权重 θ^s 更新。更新规则是 $\theta^t \leftarrow m \cdot \theta^t + (1-m) \cdot \theta^s$。动量参数 m 遵循与 BYOL（Grill et al. 2020）相同的余弦计划。在随后的实验中，我们展示了动量教师相比固定教师能带来更进一步的性能提

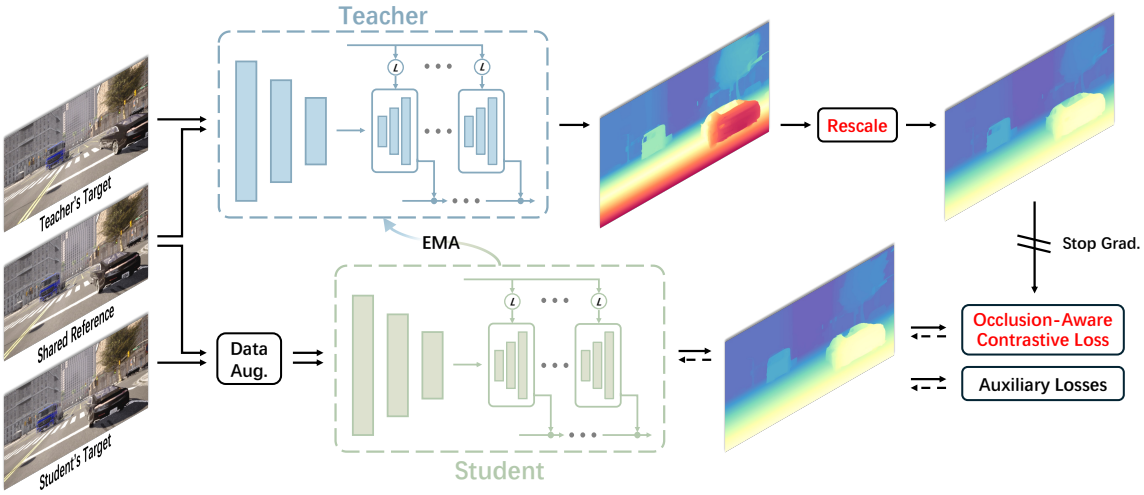

Figure 2: 我们 BaCon-Stereo 的概述。共用相同参考图像但具有不同目标图像的立体图对分别输入给教师和学生。然后教师的预测被重新调整缩放，以与学生的基线对齐。这种多视图配置使得教师能够为学生的遮挡和无遮挡区域生成可靠的伪真实值。使用数据增强，以迫使学生开发稳健的特征表示。我们对教师应用停止梯度操作，并使用学生参数的指数移动平均（EMA）更新其参数。

升。它还允许我们的框架在没有经过良好预训练的教师的情况下从头开始训练。在推理过程中，视差估计仅由教师网络执行。

遮挡感知对比损失

学生的预测 d^s 由教师的预测 d^t 与 L_1 损失监督：

$$\mathcal{L}_c = ||d^s - r \cdot d^t||_1 \tag{4}$$

，其中 $r = B^s/B^t$ 是用于将教师的基线长度调整为与学生匹配的重新缩放比率，是由 Eq. (3) 的转换得出的。

由于教师和学生有不同的目标视图，我们将每个参考图像分成三个区域：（1）对两个网络都没有遮挡的区域，（2）对教师有遮挡但对学生没有遮挡的区域，以及（3）对教师没有遮挡但对学生有遮挡的区域。在区域（2），教师经常产生较大的错误，这可能误导学生。在区域（3），学生无法仅通过匹配先验来恢复准确的视差，而教师的预测仍然可靠。此外，我们发现如 Fig. 4 所示，仅对遮挡和未遮挡区域施加统一惩罚不足以鼓励立体网络学习有效的遮挡补全。

为此，我们构建了一个遮挡感知注意力图 $\mathcal{A}$ ，该图排除区域（2）中的不可靠监督，并强调区域（3）中的监督。由于遮挡区域自然会导致较高的光度损失，我们计算教师的光度损失 $\mathcal{L}_p^t$ ，并应用一个阈值 τ 来过滤掉教师的遮挡像素。我们进一步采用自动遮蔽策略 (Godard et al. 2019) 来消除高不确定性的像素。结合这两种策略为教师产生一个二进制遮罩 $\mathcal{M}^t$ 。同样地，我们为学生生成 $\mathcal{M}^s$ 。遮挡感知注意力图 $\mathcal{A}$ 根据每个像素定义为：

$$\mathcal{A} = \begin{cases} 0, & \text{if } \mathcal{M}^t = \text{False} \\ 1, & \text{if } \mathcal{M}^t = \text{True and } \mathcal{M}^s = \text{True} \\ 2, & \text{if } \mathcal{M}^t = \text{True and } \mathcal{M}^s = \text{False} \end{cases} \tag{5}$$

通过在对比损失 $\mathcal{L}_c$ 上应用该注意力图，我们在遮挡和非遮挡区域为学生提供准确的监督，并进一步加强在教师友好但学生具有挑战性的区域的学习。

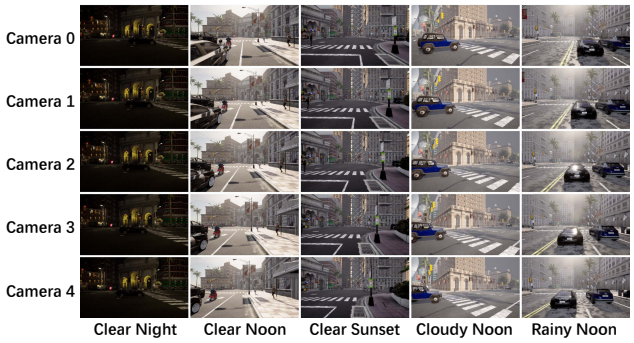

Figure 3: 我们的合成数据集 BaCon-20k 的多基线对比视差匹配示例。

使用对比损失训练的深度网络可能会陷入平凡解的崩溃（Grill et al. 2020）。对于我们的框架，教师和学生可以通过输出一个常数零视差来走捷径，这会导致 $\mathcal{L}_c$ 保持为零。为了解决这个问题，我们进一步在 Eq. (2) 和 ?? 中结合了光度损失 $\mathcal{L}_p^s$ 和平滑损失 $\mathcal{L}_s^s$ 作为辅助损失，以促进学生网络的训练。

最终损失定义如下：

$$\mathcal{L}_{total} = \mathcal{L}_c \odot \mathcal{A} + \lambda_p \mathcal{L}_p^s \odot \mathcal{M}^s + \lambda_s \mathcal{L}_s^s \tag{6}$$

其中 $\odot$ 是元素逐位相乘。参数 α 在 Eq. (2) 中被设定为 0.85，且 λ_s 被设定为 0.001 λ_p ，遵循 Monodepth2 的设置 (Godard et al. 2019) 。

合成训练数据

现有的立体匹配数据集仅提供双目图像。为了在我们的框架内训练立体网络，我们使用 CARLA 模拟器 (Dosovitskiy et al. 2017) 合成多基线立体图像。在每个时间戳，我们捕捉到五张水平偏移的彩色图像，相邻相机之间的基线一致为 0.5 米。这些图像的焦距为 480 像素，

分辨率为 540×960 像素。我们在不同的光照和天气条件下，跨越七个城镇收集了 20,417 个时间戳。 Fig. 3 显示了一些代表性的例子。

三元组输入的采样策略。在每次迭代中，我们从五张图像中随机选择一张作为参考图像，并从剩下的四张图像中采样两张作为目标图像，形成一个三元组输入。值得注意的是，我们并不限制目标图像在参考图像的右侧。当目标图像位于参考图像左侧时，我们在将其输入网络之前水平翻转这两幅图像，并相应地翻转回输出结果。

实验

数据集和评估指标

我们在我们的合成数据集 BaCon-20k 上训练立体网络，并在以下真实世界数据集上进行测试：KITTI 2015 (Menze and Geiger 2015)、KITTI 2012 (Geiger, Lenz, and Urtasun 2012)、Middlebury (Scharstein et al. 2014)、ETH3D (Schops et al. 2017) 和 DrivingStereo (Yang et al. 2019)。KITTI 2015 和 KITTI 2012 是包含数百张立体图像的驾驶场景数据集。Middlebury 是一个混合室内和室外场景的数据集，而 ETH3D 包含室内场景和灰度立体图像。Middlebury 和 ETH3D 都有几十幅图像。DrivingStereo 提供特定的天气条件（晴天、多云、雨天、雾天）。我们评估了在 KITTI 2015 和 KITTI 2012 上自监督微调的性能，KITTI 2015、KITTI 2012、Middlebury 和 ETH3D 上的合成到真实的零样本泛化性能，以及在 DrivingStereo 上对各种天气条件的鲁棒性。

我们采用 EPE（终点误差）和 k px（绝对误差大于 k 像素的离群像素百分比）作为评估指标。对于 KITTI 2015、KITTI 2012、Middlebury、ETH3D 和 DrivingStereo，离群值阈值 k 分别设置为 3、3、2、1 和 3。所有指标的值越低越好。

实施细节

我们使用 IGEVStereo (Xu et al. 2023) 作为学生和教师网络的骨干。遵循 DualNet (Wang et al. 2025)，我们从 SceneFlow (Mayer et al. 2016) 上训练的官方权重开始，并在四块 NVIDIA 4090 GPU 上继续在我们的 BaCon-20k 数据集上训练 200k 次迭代。我们使用 Adam 优化器 (Kingma and Ba 2017) 和一个周期的学习率调度 (Smith 2018)，最大学习率为 2e-4。在训练过程中，批量大小设置为 16，输入图像随机裁剪到 256×512 像素的大小。经过超参数搜索，λ_p 设置为 10，τ 设置为 0.1。

自监督微调表现

在本节中，我们评估了在 KITTI 2015 (Menze and Geiger 2015) 和 2012 (Geiger, Lenz, and Urtasun 2012) 基准上的自监督性能。

新视角外推。KITTI 2015 和 2012 只提供了双目图像对，这对于我们的框架来说是不够的。因此，我们使用生成模型在原始图像的两侧进行外推。具体来说，我们使用在 KITTI 上训练的 DEFOM-Stereo (Jiang et al. 2025) 来估算原始左右图像的视差图。接下来，我们使用估算的视差进一步向左扭曲左侧图像，并类似地向右扭曲右侧图像。然后，我们应用稳定扩散 v2 修复模型 (SDv2I) (Rombach et al. 2022) 来修复扭曲图像中遮挡的区域。在实验中，我们观察到立体网络倾向于过度平滑边缘视差，导致在扭曲的遮挡区域出现渗色伪影 (Xu et al. 2024)。这些像素应该标记为需修复的，但它们反而降低了生成外观的质量。为了缓解这个问题，我们在修复前膨胀修复掩码。最后，与之前一样，我们从这四张图像中采样图像三联体来训练立体网络。

微调性能。我们在从 KITTI 2015 和 2012 的 394 个训练对扩展出的 14,184 个三元组上微调一个周期，并提交测试集结果给基准，以便与其他自监督方法进行公平比较。如 Tab. 1 所示，我们的 BaCon-IGEV 在 KITTI 2015 和 2012 基准的所有指标上实现了最先进的自监督性能。它始终优于以前的自监督方法，例如 SsSnet (Zhong, Dai, and Li 2017) 和 PVStereo (Wang et al. 2021)。值得注意的是，SsSnet 在大规模的 KITTI 原始数据集上训练 (Geiger et al. 2013)，而 PVStereo 采用金字塔投票策略来生成用于监督的伪真值。

我们还将我们的方法与 DualNet 进行比较。在 DualNet 中，学生和教师共享相同的遮挡，阻碍教师在遮挡区域提供可靠的指导。此外，DualNet 对预测的概率分布进行监督，这限制了其与迭代方法的兼容性。相反，我们的框架利用额外的目标提示，能够在遮挡和未遮挡区域进行准确的监督，并适用于大多数立体网络。此外，动量教师使我们的框架可以从零开始训练，同时仍能达到优异的自监督性能，其效果可与甚至超越以前的最新方法相比（见 Tab. 1 的最后一行）。

零样本泛化评估

立体网络的泛化能力对于实际应用至关重要。我们评估了在场景高度不同的真实世界数据集上的零样本泛化。除非另有说明，所有比较的方法在合成数据集 SceneFlow (Mayer et al. 2016) 上进行训练，并在真实世界数据集的训练集上进行评估 (Menze and Geiger 2015; Geiger, Lenz, and Urtasun 2012; Scharstein et al. 2014; Schops et al. 2017)。

如 Tab. 2 所示，通过将遮挡注意力对比损失应用于 IGEVStereo (Xu et al. 2023)，我们的方法在遮挡区域平均减少了 4.67 % 的异常值，同时大大提高了非遮挡区域的预测准确性。我们的 BaCon-IGEV 在所有指标中均排名前三，其性能与 DEFOMStereo-L (Jiang et al. 2025) 相当，但参数减少了 30 × 。此外，我们的方法在遮挡和非遮挡区域均优于 NerfStereo (Tosi et al. 2023)。尽管 NerfStereo 在反向传播期间利用了来自非遮挡第三视图的光度损失，但这一设计主要增强了匹配性能，而未显著促进遮挡补全。这些结果凸显了我们的 BaCon-Stereo 框架在引导立体网络学习非遮挡区域的精确匹配和遮挡区域的有效补全方面的效能。

如 Fig. 4 所示，IGEVStereo (Xu et al. 2023) 和 Selective-IGEV (Wang et al. 2024) 都对遮挡和未遮挡区域施加了统一的监督强度，这常常导致遮挡区域的预测不准确。相比之下，我们的 BaCon-IGEV 显著改善了这一问题，得益于由遮挡感知注意力图指导的对比损失。

跨不同天气的稳健性

对于在室外操作的立体视觉系统来说，在恶劣天气条件下的鲁棒性是必不可少的。Tab. 3 表明我们的方法在 DrivingStereo (Yang et al. 2019) 上实现了卓越的跨天气鲁棒性。相比于 IGEVStereo，我们将晴天、阴天、雨天和雾天情况下的异常值比例减少到了至少原来的一

Method	KITTI 2015						KITTI 2012					
	NOC			ALL			NOC			ALL		
	D1-BG	D1-FG	D1-ALL	D1-BG	D1-FG	D1-ALL	EPE	>2px	>3px	EPE	>2px	>3px
SsSnet (arXiv 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
MC-CNN-WS (ICCV 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
OASM (ACCV 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 (CVPR 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 (ECCV 2020)	2.97	8.33	3.86	3.13	8.70	4.06	—	—	—	—	—	—
PVStereo (RA-L 2021)	2.09	5.73	2.69	2.29	6.50	2.99	0.7	4.55	1.98	0.8	5.25	2.47
PASMnet (T-PAMI 2022)	5.02	15.16	6.69	5.41	16.36	7.23	1.3	—	7.14	1.5	—	8.57
EMR-MSF (ICCV 2023)	8.30	14.16	9.27	8.61	15.15	9.70	—	—	—	—	—	—
UHP (RA-L 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
DualNet (AAAI 2025)	2.28	4.66	2.67	2.46	5.25	2.92	0.6	—	2.06	0.6	—	2.59
BaCon-IGEV (Ours)	2.06	4.43	2.45	2.21	4.86	2.65	0.5	3.11	1.92	0.6	3.69	2.31
BaCon-IGEV † (Ours)	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 1: 在 KITTI 2015 和 2012 基准上的定量结果。† 从头开始训练的 BaCon-IGEV。KITTI 2015 和 2012 评估在非遮挡（NOC）和所有（ALL）区域中的性能。KITTI 2015 进一步报告了 D1 指标（即绝对误差大于 3 像素且相对误差大于 5 % 的异常值比例），并在背景（BG）、前景（FG）和所有（ALL）区域中进行报告。第一 、第二 和 第三 分别由颜色表示。

Method	KITTI 2015			KITTI 2012			Middlebury			ETH3D			# Params.
	OCC	NOC	ALL	OCC	NOC	ALL	OCC	NOC	ALL	OCC	NOC	ALL	(M)
PSMNet (CVPR 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	5.22
GwcNet (CVPR 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	6.91
CFNet (CVPR 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	23.05
RAFTStereo (3DV 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	11.12
ACVNet (CVPR 2022)	32.85	11.29	11.71	54.47	12.94	13.89	47.36	22.07	25.66	19.64	8.65	9.19	7.17
PCWNet (ECCV 2022)	14.95	5.53	5.74	30.22	4.07	4.67	37.99	12.17	15.86	11.67	5.28	5.54	35.94
IGEVStereo (CVPR 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	12.60
NerfStereo-RAFT † (CVPR 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	11.12
SelectiveIGEV (CVPR 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	13.14
DEFOMStereo-L ‡ (CVPR 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	382.62
BaCon-IGEV (Ours)	10.37	4.05	4.21	21.49	3.08	3.48	23.46	6.34	8.94	7.98	2.51	2.76	12.60

Table 2: 跨真实世界数据集的零样本泛化。我们报告遮挡区（OCC）、非遮挡区（NOC）和所有（ALL）区域的异常值指标。所有比较的方法都使用其官方发布的权重进行重新评估，以确保一致的基准测试。† 在具有伪真实值的真实世界数据上训练的方法。‡ 基于视觉基础模型的方法。

半。值得注意的是，在所有竞争方法都产生超过 10 % 异常值并因此对基于立体视觉的系统构成严重风险的雨天条件下，我们的方法仅以 5.07 % 异常值保持了稳健的准确性。这些成果主要归功于我们的整体框架，它结合了学生的数据增强与教师的动量更新策略以促进有效学习。学生必须学习强大的特征表示能力，以及稳健的匹配和完成能力，以应对开放世界。因此，即使在训练过程中没有雾天样本，我们的方法在雾天条件下仍表现出强大的性能。

Fig. 5 展示了在雨雾条件下的定性结果。具体来说，IGEVStereo 和 Selective-IGEV 都在积水路面上失败，而我们的方法仍然能预测准确的视差。

我们通过全面的消融实验来验证我们方法的有效性。所有变体都在我们的 BaCon-20k 数据集上进行训练，并报告在 KITTI 2015 训练集上的遮挡 (OCC)、非遮挡 (NOC) 和所有 (ALL) 区域的异常值 (>3px) (Menze and Geiger 2015) 。

损失项。如 Tab. 4 所示，传统的自监督损失 (A) 无法为被遮挡区域提供有效监督，导致几乎完全错误的预测。这一局限性源于光度监督在被遮挡区域中无法提供可靠的学习信号，在这些区域立体图像之间不存在真正的对应关系。如果处理不当，模型往往会过拟合错误的光度线索，导致在这些具有挑战性的区域中出现错误的视差预测。相反，我们的对比损失 (B) 通过利用多基线几何一致性解决了这个问题，实现了在被遮挡和未被遮挡区域的同时性能提升。我们进一步研究了这些损失函数 (C)-(E) 的组合。定量结果证实，综合损失 (E) 提供了最佳的结果。

动量教师。在 Tab. 4 中，我们进一步证明了动量教师的好处。与固定教师 (E) * 相比，动量教师 (E) 提供了更稳定和更具普遍性的监督，导致在所有区域都实现了一致的改进。

遮挡感知注意力图。我们对生成二值掩膜的两种策略进行了消融测试。如 Tab. 5 所示，自动掩膜策略 (Godard et al. 2019) 通过去除高不确定性像素（如 i.e. 、纹理较弱的区域和无穷远区域）有效地提高了性能。在此基础上，过滤掉遮挡的像素（这些像素固有地表现出显著的光度损失）进一步提升了掩膜的质量，减少了监督信号中的噪声，从而实现了更优的视差预测。通过对遮挡区域施加更强的监督，立体网络在这些具有挑战

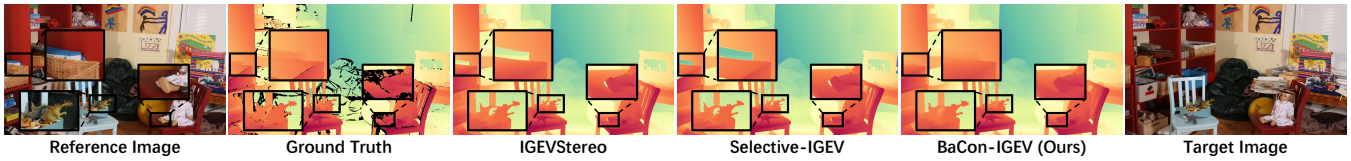

Figure 4: Middlebury (Scharstein et al. 2014) 数据集中遮挡区域中对 IGEVStereo (Xu et al. 2023)、Selective-IGEV (Wang et al. 2024) 和 BaCon-IGEV（我们的工作）的定性比较。

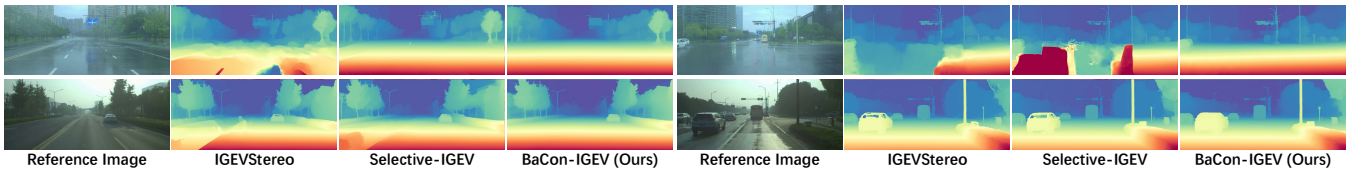

Figure 5: 在 DrivingStereo (Yang et al. 2019) 的雨天（上排）和雾天（下排）子集上对 IGEVStereo (Xu et al. 2023)、Selective-IGEV (Wang et al. 2024) 和 BaCon-IGEV（我们的）进行定性比较。

Method	Sunny	Cloudy	Rainy	Foggy
PSMNet (2018)	40.14	43.95	56.19	69.69
GwcNet (2019)	17.12	25.56	28.19	29.23
CFNet (2021)	4.70	5.30	12.48	5.54
RAFTStereo (2021)	4.23	4.19	12.77	3.09
ACVNet (2022)	19.64	29.87	41.31	38.05
PCWNet (2022)	3.58	3.67	10.52	5.20
IGEVStereo (2023)	4.59	5.15	15.47	4.49
NerfStereo † (2023)	2.88	2.91	10.20	3.93
Selective-IGEV (2024)	5.05	5.24	13.51	4.10
DEFOMStereo ‡ (2025)	3.61	3.75	13.53	2.88
BaCon-IGEV (Ours)	2.15	1.87	5.07	1.64

Table 3: 驾驶立体视觉在不同天气条件下的稳健性 (Yang et al. 2019)。报告离群值（>3px）。† 该方法在真实世界数据上训练，并使用伪真实值。‡ 基于视觉基础模型的方法。

	$\mathcal{L}_c$	$\mathcal{L}_p$	$\mathcal{L}_s$	OCC	NOC	ALL
(A)		✓	✓	97.44	5.51	7.14
(B)	✓			11.22	5.23	5.38
(C)	✓	✓		10.62	4.49	4.64
(D)	✓		✓	11.23	5.14	5.29
(E)	✓	✓	✓	10.37	4.05	4.21
(E) *	✓	✓	✓	10.87	4.98	5.12

Table 4: 在 Eq. (6) 中对损失项进行消融研究。* EMA 不用于更新教师网络。

Threshold	Auto-Mask	Occ-Aware	OCC	NOC	ALL
			13.67	4.93	5.13
✓			13.49	4.76	4.96
	✓		13.27	4.43	4.64
✓	✓		12.39	4.06	4.25
✓	✓	✓	10.37	4.05	4.21

Table 5: 有效遮罩和考虑遮挡的注意力图的消融研究。

性的区域取得了显著改进，证明了我们的遮挡感知机制的有效性。

通用性。我们的对比损失直接监督预测的视差图，使得我们的框架适用于各种类型的立体视觉网络。我们在基于 3D 卷积的骨干网络 (Guo et al. 2019; Shen, Dai, and Rao 2021) 和基于迭代的骨干网络 (Lipson, Teed, and Deng 2021) 上评估其通用性。如 Tab. 6 所示，我们的方法一致地提升了所有三个基线的性能。尽管 RAFTStereo 在遮挡区域已经表现良好，我们的方法进一步将其完成能力从 12.70 % 提升至 11.73 % 。

结论

在本文中，我们提出了 BaCon-Stereo，这是一种通过多基线对比学习训练的自监督师生框架。通过利用教师和学生之间不同的非遮挡区域，我们为学生提供了可靠的监督，覆盖了遮挡和非遮挡区域。此外，我们引入了一个遮挡感知注意力图来指导学生学习遮挡区域的精确视差补全。为了验证我们方法的有效性，我们构建了 BaCon-20k，这是一个涵盖多种光照和天气条件的多基线立体数据集。实验结果表明，我们的方法在不同场景广泛化和在不同天气条件下的鲁棒性方面能与现有方法匹敌或超越。我们还在 KITTI 2015 和 2012 基准测试中优于之前的自监督方法。

References

Aleotti, F.; Tosi, F.; Zhang, L.; Poggi, M.; and Mattoccia, S. 2020. Reversing the cycle: self-supervised deep stereo through enhanced monocular distillation. In European Conference on Computer Vision, 614–632. Springer.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In International conference on machine learning, 233–242. PMLR.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, 9650–9660.

Method	OCC	NOC	ALL
GwcNet (CVPR 2019)	29.07	12.17	12.53
BaCon-Gwc (Ours)	12.70	4.46	4.65
CFNet (CVPR 2021)	16.42	5.87	6.10
BaCon-CF (Ours)	11.88	4.26	4.44
RAFTStereo (3DV 2021)	12.70	5.34	5.53
BaCon-RAFT (Ours)	11.73	3.92	4.09

Table 6: 我们的框架通用性。我们另外测试了基于3D 卷积的骨干网络 GwcNet (Guo et al. 2019) 和 CFNet (Shen, Dai, and Rao 2021)，以及基于迭代的骨干网络 RAFTStereo (Lipson, Teed, and Deng 2021)。

Chang, J.-R.; and Chen, Y.-S. 2018. Pyramid stereo matching network. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5410–5418.

Chen, S.; Xiang, Z.; Xu, P.; and Zhao, X. 2022. A normalized disparity loss for stereo matching networks. IEEE Robotics and Automation Letters, 8(1): 33–40.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In Conference on robot learning, 1–16. PMLR.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets Robotics: The KITTI Dataset. International Journal of Robotics Research (IJRR).

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition, 3354–3361. IEEE.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF international conference on computer vision, 3828–3838.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33: 21271–21284.

Guo, X.; Yang, K.; Yang, W.; Wang, X.; and Li, H. 2019. Group-wise correlation stereo network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3273–3282.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 9729–9738.

Jiang, H.; Lou, Z.; Ding, L.; Xu, R.; Tan, M.; Jiang, W.; and Huang, R. 2025. DEFOM-Stereo: Depth Foundation Model Based Stereo Matching. arXiv preprint arXiv:2501.09466.

Jiang, Z.; and Okutomi, M. 2023. EMR-MSF: Self-Supervised Recurrent Monocular Scene Flow Exploiting Ego-Motion Rigidity. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 69–78.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Li, A.; and Yuan, Z. 2018. Occlusion aware stereo matching via cooperative unsupervised learning. In Asian Conference on Computer Vision, 197–213. Springer.

Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; and Liu, S. 2022. Practical stereo matching via cascaded recurrent network with adaptive correlation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16263–16272.

Lipson, L.; Teed, Z.; and Deng, J. 2021. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In 2021 International Conference on 3D Vision (3DV), 218–227. IEEE.

Liu, P.; King, I.; Lyu, M. R.; and Xu, J. 2020. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 6648–6657.

Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; and Brox, T. 2016. 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, 4040–4048.

Menze, M.; and Geiger, A. 2015. Object scene flow for autonomous vehicles. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3061–3070.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10684–10695.

Scharstein, D.; Hirschmüller, H.; Kitajima, Y.; Krathwohl, G.; Nešić, N.; Wang, X.; and Westling, P. 2014. High-resolution stereo datasets with subpixel-accurate ground truth. In German conference on pattern recognition, 31–42. Springer.

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

Schops, T.; Schonberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. 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, 3260–3269.

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

Shen, Z.; Dai, Y.; et al. 2022. PCW-Net: Pyramid Combination and Warping Cost Volume for Stereo Matching. In ECCV.

Smith, L. N. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv:1803.09820.

Tosi, F.; Tonioni, A.; De Gregorio, D.; and Poggi, M. 2023. Nerf-supervised deep stereo. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 855–866.

Tulyakov, S.; Ivanov, A.; and Fleuret, F. 2017. Weakly supervised learning of deep metrics for stereo reconstruction. In Proceedings of the IEEE International Conference on Computer Vision, 1339–1348.

Wang, H.; Fan, R.; Cai, P.; and Liu, M. 2021. PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching. IEEE Robotics and Automation Letters, 6(3): 4353–4360.

Wang, L.; Guo, Y.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; and An, W. 2022. Parallax Attention for Unsupervised Stereo Correspondence Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(4): 2108–2125.

Wang, X.; Xu, G.; Jia, H.; and Yang, X. 2024. Selective-stereo: Adaptive frequency information selection for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19701–19710.

Wang, Y.; Zheng, J.; Zhang, C.; Zhang, Z.; Li, K.; Zhang, Y.; and Hu, J. 2025. DualNet: Robust Self-Supervised Stereo Matching with Pseudo-Label Supervision. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 8178–8186.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4): 600–612.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3733–3742.

Xu, G.; Cheng, J.; Guo, P.; and Yang, X. 2022. Attention Concatenation Volume for Accurate and Efficient Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12981–12990.

Xu, G.; Wang, X.; Ding, X.; and Yang, X. 2023. Iterative Geometry Encoding Volume for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 21919–21928.

Xu, P.; Xiang, Z.; Fu, J.; Pu, T.; Zhong, H.; and Liu, E. 2025. MIDAS: Modeling Ground-Truth Distributions with Dark Knowledge for Domain Generalized Stereo Matching. arXiv preprint arXiv:2503.04376.

Xu, P.; Xiang, Z.; Qiao, C.; Fu, J.; and Pu, T. 2024. Adaptive Multi-Modal Cross-Entropy Loss for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5135–5144.

Yang, G.; Song, X.; Huang, C.; Deng, Z.; Shi, J.; and Zhou, B. 2019. 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, 899–908.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10371–10381.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything v2. Advances in Neural Information Processing Systems, 37: 21875–21911.

Yang, R.; Li, X.; Cong, R.; and Du, J. 2024c. Unsupervised hierarchical iterative tile refinement network with 3D planar segmentation loss. IEEE Robotics and Automation Letters, 9(3): 2678–2685.

Žbontar, J.; and LeCun, Y. 2016. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 17(65): 1–32.

Zhong, Y.; Dai, Y.; and Li, H. 2017. Self-supervised learning for stereo matching with self-improving ability. arXiv preprint arXiv:1709.00930.