视窗动态模糊（WinDB）：无头戴式显示器（HMD）且无失真的全景视频注视点学习方法

王国涛, 陈程立诏, 郝爱民, 秦洪, 范登平

摘要—截至目前，在全景视频领域，被广泛采用的注视点采集方式是借助头戴式显示器（HMD）来完成的。具体而言，用户通过佩戴HMD自由地对给定的全景场景展开探索，在此过程中同步完成注视点的收集工作。然而，这种常见的数据采集方式在用于训练深度模型时存在明显缺陷，它难以精准判定在给定的全景视频存在间歇性显著事件的情况下，究竟哪些区域才是最为关键的。其主要原因在于，使用HMD采集注视点时必然会出现“视野盲区”。毕竟，用户不可能一刻不停地转动头部去全方位探索整个全景场景。如此一来，所采集到的注视点往往仅集中于某些局部视角，而其他区域便沦为未被关注到的“视野盲区”。因此，依靠基于HMD的方法所收集与积累的局部视图注视点数据，无法准确呈现出复杂全景场景的整体全局重要性，而这恰恰是注视点采集的关键要点所在。为有效解决这一难题，本文提出了一种应用于全景视频的带有动态模糊效果的辅助窗口（WinDB）注视点采集方法。该方法无需头戴式显示器，且能够出色地体现出区域层面的重要程度。凭借WinDB方法，我们推出了全新的数据集（PanopticVideo-300），此数据集涵盖300个全景视频片段，涉及超过225个类别。尤其值得关注的是，由于WinDB方法在采集注视点时不存在“视野盲区”的问题，所以在我们的新数据集中呈现出一种极为特殊且在以往研究中长期被忽视的现象——频繁且密集的“注视点转移”。针对这一情况，我们专门设计了一种高效的注视点转移网络（FishNet）来加以处理。所有这些全新的注视点采集工具、数据集以及网络，极有希望为360°环境下注视点相关的研究与应用开启崭新的篇章。

Index Terms—无头戴式显示器（HMD），无失真，全景视频注视点学习

1 介绍和动机

对于给定的全景场景而言，全景注视点预测的核心目标在于感知区域层面的重要性，也就是要反映出整个场景里不同区域所获得的不同关注程度。通过达成这一目标，能够实现对场景中最为“重要区域”的快速定位。

一般而言，定位到的重要区域有着诸多广泛的应用。具体来讲，正如图 1-D所展示的那样，全景视频导航 [1]有助于在盲区时定位间歇性出现的“重要区域”。

与已经被广泛研究的传统2D注视点预测方法 [6]–[9]有所不同，全景注视点预测 [1], [10]–[14]目前仍处于起步阶段。导致其进展缓慢的主要问题是缺少大规模的数据集 [10], [15], [16]，毕竟在全景场景中采集人眼注视点要比在传统2D场景中困难得多 [17]–[20]。此外，全景注视点预测比传统的2D图像注视点预测更为复杂，2D数据仅有一个固定视角，而全景数据却能让用户自由探索360°的全景视频 [21]–[23]。所以，当前的全景注视点预测研究领域正面临一个两难的局面——要用非常小规模的训练数据[1]去解决一个复杂的问题。

截至目前，基于头戴式显示器（HMD）的人眼注视点采集方法 [2]–[5]是最为常用的方法。用户通过佩戴HMD自由探索给定的全景场景，与此同时采集注视点数据 [24]–[26]。尽管应用广泛，但是HMD注视点采集方法 [2]–[5]存在两个问题，其中一个尤为关键。其一，使用HMD采集注视点时，总会出现“盲区”（blind zoom）的问题，这是因为用户不可能始终不停地转动头部去探索整个全景场景。盲区问题使得HMD采集的注视点与全景场景中实际的区域重要性程度不匹配。所以，在“盲区”内发生的显著事件可能完全没有注视点（详细内容见 Sec.6.1.4）。为了能更好地理解这一点，我们在图1-A和B中给出了一个形象的示例。其二，HMD注视点采集方法相对“昂贵”[2]，并且用户佩戴HMD探索全景场景时通常会感觉极为不适（例如，VR晕动症 [27]、视觉疲劳 [28]、重量和平衡问题 [29]，详情见图 1-C-a）。简而言之，基于HMD的注视点采集方法在方法上存在缺陷，其标注过程也相当昂贵。

除了前面提到的基于头戴式显示器（HMD）的注视点预测方法，值得一提的是，基于普通等距矩形投影（Equi-Rectangular Projection, ERP）的方法 [30]–[35]，即将全景场景这一典型的球形数据投影到2D平面上，通常也不适合作为人眼注视点采集的平台。主要原因在于，人眼视觉系统对于发生在具有强烈ERP视觉畸变区域中的显著或重要事件的敏感度比较低 [36]–[40]。因此，通过ERP投影方式收集的人眼注视点无法很好地反映给定全景场景中的区域重要性水平。具体有两个原因。第一，如图 2-ERP所示，靠近ERP极点的强烈视觉畸变会给人眼视觉系统带来较大负担，使得用户很

- 本文为*TPAMI*论文的中文翻译版本。
- 通讯作者：陈程立诏（*cclz123@163.com*）。

[1]常用的208段视频片段，大多是简单场景，无法满足复杂模型的训练需求。

[2]这里的昂贵主要指的是标注成本，而非设备成本。

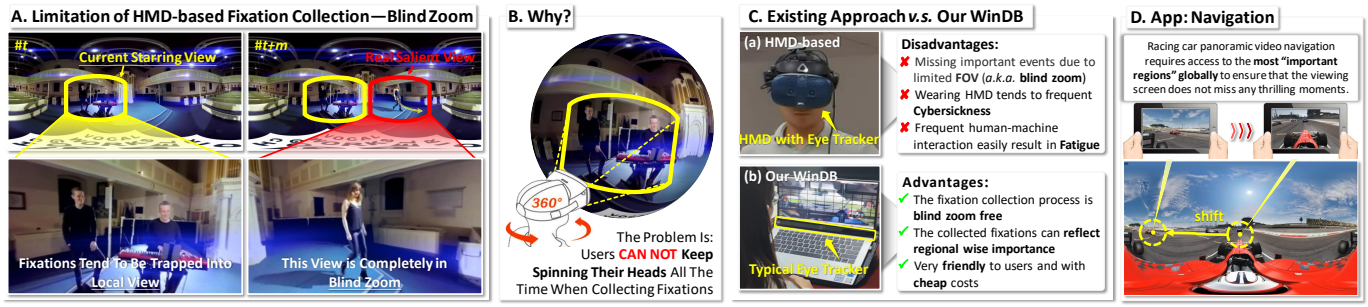

图 1. 基于头戴式显示器（HMD）的全景数据注视点采集方法 [2]–[5]存在一个关键限制——盲区（blind zoom）。这一限制致使采集到的注视点不足以训练深度模型，从而无法准确预测给定全景场景中哪些区域最为重要（A）。造成这一问题的原因如（B）所示：佩戴HMD的用户在场景探索的初期阶段后，往往会变得“迟缓”（retard），进而导致错过在盲区中发生的重要事件。（C）总结了我们提出的新型WinDB方法相较于现有基于HMD的注视点采集方法的优势，其中优势和劣势分别用 ✓和 ✗表示。详细内容请参见 Sec. 1。（D）能够充分反映场景区域重要性的全景注视点可应用于多种场景，例如全景视频导航 [1]，该应用能够帮助定位盲区中间歇性出现的“显著事件”。其中 *t*和 *t+m*分别表示第 *t*帧和第 *t+m*帧，*m*为1到15之间的随机值。由于人类视觉系统（HVS）的响应极限为500毫秒，这相当于大约15帧。

难判断是否存在任何异常或重要事件值得进一步关注。所以，在ERP极点附近经常会出现毫无意义的快速扫视注视点（见中间的示例）。第二，由于上述原因，长时间的全景场景探索会导致视觉系统疲劳，这样视觉系统可能在长时间注视后变得迟钝，从而错过那些发生在畸变ERP区域中的重要事件 [41]–[44]（见右侧的示例）。此外，使用多个鱼眼图像 [45]–[49]，这些图像具有较大的视场，也面临着类似于ERP方法的问题，比如盲区和视觉畸变之间的权衡。360°的视场会引入显著的边缘畸变，而90°的视场会导致拓扑信息的丢失，多个鱼眼图像中的重叠视场则会引发重影效应。

鉴于上述种种情况，本论文提出了一种新颖的方法，叫做视窗动态模糊（WinDB），用于全景注视点采集。我们的WinDB方法充分考量了基于HMD和基于ERP的方法的优缺点。其核心思想是充分利用ERP方法的优点（即无盲区），并通过一系列专门设计来抑制视觉畸变（见图 1-C-b）。我们可以在图 4中预览其总体框架。通过WinDB方法采集的注视点能够很好地指示给定全景场景中每个区域的重要性程度（见图 2-WinDB）。

此外，利用我们提出的WinDB方法，我们构建了一个大型的全景视频注视点预测数据集——PanopticVideo-300，这是360°视频注视点预测领域中最具挑战性的数据集，因为它显著包含了盲区场景。通过解决盲区问题，WinDB采集的注视点能够有效地反映区域的重要性，使得PanopticVideo-300成为我们研究领域中的第一个综合性数据集。

图 2. ERP与WinDB在视觉畸变去除和注视点采集中的比较。 ERP的视觉畸变会导致无意义的注视点，并随着时间的推移错过重要事件。而WinDB由于没有畸变，能够实现准确且有意义的注视点采集。

关于我们的PanopticVideo-300，我们发现了一个有趣的现象——频繁的“注视点转移”。由于HMD方法的局限性（即盲区），通过HMD采集的注视点在传统的全景注视点数据集中通常显得较为平滑（例如，VR-EyeTracking [4]和Wild360 [42]）。正如我们所提到的，由于盲区，HMD采集的注视点往往局限于局部视角，所以这些注视点通常是平滑的。与之形成鲜明对比的是，我们的WinDB方法不存在盲区问题；因此，那些原本会被HMD方法忽略的显著事件现在能够被完整地捕捉到。所以，在我们的PanopticVideo-300数据集中，注视点可能会在极短的时间内迅速转移到另一个距离较远的位置（见图 3-A），我们将这种现象称为“注视点转移”。这种现象并非坏事；相反，它进一步证实了我们WinDB方法在注视点采集中的可靠性。为了能更好地理解，我们在图 3-B中给出了一个生动的例子，展示了注视点从正在讲话的人转移到推门进来的男子身上的过程。WinDB的相关技术细节和更深入的分析将在 Sec.3中给出。进一步来说，我们面临另一个困境——即之前的任何注视点预测网络都无法很好地处理“注视点转移”现象（有关详细信息，见 Sec.6.2.2和 Sec.6.2.3）。造成这种不足的主要原因通常有两个：1) 这些网络的设计过度注重预测注视点的时空平滑性；2) 它们对注视点的感知范围基本上是局部的，这使得从理论上讲不可能感知到长距离的注视点转移。因此，本文还提出了一种新的网络（即FishNet）来处理注视点转移现象，其技术原理极具启发性，并且有潜力为未来的研究提供指导。关于这一新颖内容的详细讨论将在 Sec.5中展开。

总结来说，本文的主要贡献包括以下几点：

- 我们提出了一种全新的注视点采集方法（**WinDB**），该方法首次突破了由头戴式显示器（HMD）引发的盲区限制。
- 依托WinDB，我们推出了一个全新的数据集——Panoptic Video-300，这是首个坚实且极具挑战性的360°视频注视点预测数据集，其注视点能够切实反映区域的重要性。
- 作为开创性的首次尝试，我们设计了一种全新的网络设计范式，以应对“注视点转移”挑战，并将这个新网络命名

图 3. 定性展示了我们WinDB方法与HMD方法（即VR眼动追踪 [4]）采集的数据集之间的差异。子图A展示了在PanopticVideo-300 数据集中出现的注视点转移现象。由于我们的WinDB方法没有盲区，它能够捕捉到HMD方法忽略的显著事件，即发生在盲区内的人类和动物突发事件。子图B展示了“注视点转移”现象，其中注视点从正在讲话的人转移到推门的人。然而，由于HMD方法存在盲区，它的注视点集中在了正在讲话的人身上。

为FishNet。

- 本文开展了一整套工作，涵盖数据采集方法（WinDB[3]）、网络（FishNet[4]）以及数据集（PanopticVideo-300[5]），其方法论、最新研究成果、深入剖析以及结论将共同推动我们所在研究领域的发展进程。

2 相关工作

2.1 注视点采集方法

基于HMD的方法：基于HMD（头戴式显示器）的方法 [2]–[5]要求用户佩戴头戴式显示器，借助其内置传感器来采集注视点数据。因为用户的视角与头部运动能够自然地保持同步，所以这种方法能够有效消除畸变。不过，用户只能看到HMD视场内的内容，这样就有可能会遗漏视场外的重要信息。此外，HMD方法必须依靠专用硬件的支持，这在一定程度上限制了它的扩展性与可访问性。

基于ERP的方法：基于ERP（等距矩形投影）的方法 [42], [50]是在计算机屏幕上展示全景视频内容，通过眼动追踪技术 [42]或者鼠标输入 [50]来采集注视点数据。这些方法允许用户不受固定视角的限制，可以自由地探索整个全景场景。然而，将全景视频呈现在矩形屏幕上会引入视觉畸变，进而对注视点数据的准确性产生不利影响。尽管如此，基于ERP的方法相对来说更容易获取，因为它们除了普通眼动追踪系统或鼠标之外，不需要其他专用硬件。

总之，基于HMD的方法虽然能够提供沉浸式的视角体验，但存在视场受限的弊端；而基于ERP的方法虽然没有盲区的限制，但却会引入畸变问题。所以，我们迫切需要一种能够同时消除盲区和畸变的有效解决方案。

2.2 注视点学习网络

全景注视点学习旨在对全景视频中的重要区域进行预测，相较于传统的2D方法，其复杂程度更高，这一点在 [6]–[9]等文献中均有提及。当前的方法大致可分为以下三大类：

双流融合网络： 这类网络会整合全局ERP和局部CMP信息，采用诸如动态加权融合 [51]以及双投影融合 [52]等技术手段。然而，ERP和CMP特征之间的对齐问题可能会导致最终结果不太理想。

[3]https://github.com/guotaowang/WinDB.
[4]https://github.com/guotaowang/FishNet.
[5]https://github.com/guotaowang/PanopticVideo-300.

基于球面卷积的网络： [2], [10], [14], [53]–[55]所介绍的这些方法，是把ERP坐标映射到球面以开展卷积操作，从而维持空间关系。但是，这些方法无法使用经过预训练的2D特征骨干网络，而预训练网络对于提取丰富的语义信息起着至关重要的作用，所以它们的性能会受到一定程度的限制。

基于Transformer的网络： [56], [57]中的这些网络，运用可变形卷积将球面数据映射为2D补丁，以此减少ERP畸变。虽然这种方法在特征对齐方面有所改善，但中间特征仍可能存在部分畸变，这就限制了预训练网络参数在2D图像中的有效应用。

总之，尽管全景注视点学习已经取得了显著的进展，但当前的方法在融合全局与局部信息、处理畸变以及使用预训练模型等方面依旧面临诸多挑战。我们的模型致力于通过提出一个全新的框架来攻克这些局限性，该框架具备全局性、无畸变性以及语义强度等优势，这些特性使得它在性能表现上能够超越上述各类方法。

3 新型360°注视点采集方法（WinDB）

3.1 WinDB概述

为了充分发挥ERP（等距矩形投影）以及HMD（头戴式显示器）方法的优势，特别是在消除盲区这一方面，WinDB选取ERP作为注视点采集的基础计算平台。这样一来，关键之处就在于要尽可能地消除因畸变而带来的负面影响。我们在图 4中给出了WinDB的整体概览，它包含有五个步骤。

1. 解决畸变问题： 我们采用了步骤A——网格状球面到2D的投影（具体可参照图 4-**a**），此步骤会把输入的ERP图像均匀划分成多个子区域。随后，将每个球面子区域投影到平面的2D图像块上。因此，当步骤A完成之后，所有的局部子图块都不存在畸变了，大家可查看图 4中标记为②的黄色部分。不过，即便如此，我们仍然能够明显地察觉到“鬼影效应”以及大量存在的“子图块之间的不对齐”现象（具体可查看图 4-**a**中的红色箭头所指之处）。显然，在进行注视点采集的时候，这些伪影是不被期望出现的，因为它们往往会吸引人们的目光。

2. 改进方案： 为了对上述情况加以改进，我们精心设计了一个十分巧妙的步骤（也就是图 4中所标记的步骤B）——我们使用黑色窄线来遮挡相邻子图块的拼接，通过这样的操作使得整个图像都被覆盖上了黑色的“网格屏幕”。这个步骤的设计灵感源自于我们人类视觉系统所具有的一个独特机制，

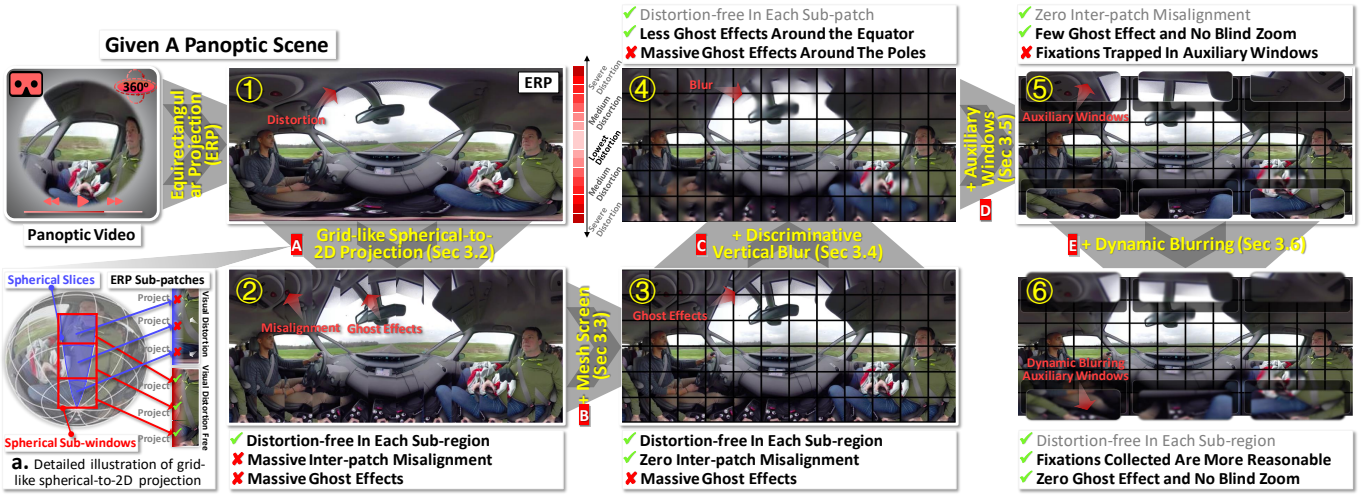

图 4. 我们新提出的无HMD注视点采集方法的整体流程。与广泛使用的基于HMD的方法相比，我们的WinDB方法在经济性、舒适性和合理性方面具有优势。在给定的全景视频场景中，最**严重**的畸变将通过**D**和**E**来解决，而**中等**畸变则通过**A**、**B**和**C**来解决。详细信息见Sec.3。

即“视觉暂留”（Persistence of Vision，POV，相关内容可参考 [49], [58], [59]，其示意图可参照图 5），这意味着我们的视觉系统能够自动对被遮挡的窄区域进行恢复。所以，通过运用这种设置“网格屏幕”的巧妙方式，我们就能够轻松地处理“子图块之间的不对齐”问题了（具体可查看图 4中标记为③的黄色部分）。关于更多的技术细节，将会在Sec. 3.3章节中给出详细介绍。

3. 处理“鬼影效应”： 接下来，我们需要对“鬼影效应”进行处理。在此，我们提出了一个既简单又行之有效的解决方案（即图 4中所标记的步骤C），也就是要在每个局部图块上执行“判别性垂直模糊”操作[6]。而且，模糊的程度会依据图块所处的位置来进行动态调整，对于那些靠近极点的图块，其模糊程度会被设置得更厉害一些。通过采取这个步骤，能够使鬼影效应得到显著的减轻；大家可查看图 4中标记为④的黄色部分以了解效果。关于更多的技术细节，将会在Sec. 3.4章节中给出详细介绍。

4. 进一步解决鬼影效应： 然而，有一些图块，尤其是那些靠近极点的图块（也就是位于顶部和底部的行），仍然存在着较为明显的鬼影效应。所以，我们又提出了一个新的步骤（即图 4中所标记的步骤D），那就是在极点周围设置“辅助窗口”。这些辅助窗口从本质上来说，其实就是尺寸相对较大的子图块，并且它们是不存在畸变的。通过合理地对窗口覆盖率进行分配，就能够彻底地解决“鬼影效应”问题；大家可查看图 4中标记为⑤的黄色部分以了解效果。更多技术细节将在 Sec. 3.5中给出。

5. 动态模糊辅助窗口： 此外，由于这些辅助窗口不存在畸变，而且相较于局部图块而言，它们还包含有更多的信息，所以用户自然会更加倾向于去关注它们。为了确保注视点采集过程的客观性，我们采取了步骤E——动态模糊，也就是要逐步对那些已经接收到人眼注视的辅助窗口进行模糊处理。关于更多的技术细节，将会在Sec. 3.6章节中给出详细介绍。

上述所有的步骤共同构成了我们所提出的WinDB方法，该方法具有无畸变、无鬼影效应、无视觉伪影、无盲区等优点，并且对用户来说也是十分友好的。

3.2 网格状球面向二维投影

通常情况下，采用ERP来表示全景图像能够保留图像的整体信息。不过，正如我们在图 4-①中所展示的那样，ERP图像中是存在视觉畸变的，并且在靠近极点的地方，这种畸变程度还会变得愈发严重。

受到人类视觉系统（HVS）焦距范围有限这一特性的启发[7]，我们提出了“网格状球面到2D投影”（具体可参照图 4-**a**）的方法，通过这种方式来让ERP达成“局部无畸变”的效果，也就是要使得每个子图块都不存在畸变。尽管在采用这种“网格状球面到2D投影”的过程中，在垂直方向上不可避免地会引入一些较为明显的子图块间错位情况，但我们认为，倘若能够达成“**前提条件**”——也就是让我们的HVS能够聚焦于子图块内部区域，那么经过投影之后的ERP也可以被视作是无畸变的。

在此处，我们先将如何达成这个“前提条件”的问题暂且留待后续的小节去进行讨论，接下来就开始详细阐述我们所提出的“网格状球面到2D投影”这一方法。

如图 4-**a**所示，我们先假定有一个典型的ERP帧被投影到了球面上；然后，我们会将这个球面划分成多个球面切片，在垂直方向上，我们将划分间隔设定为30° [8]。然而，在水平方向上，其间隔是动态变化的，这样做的目的在于确保所得到的球面切片能够具备与ERP网格相同的网格拓扑结构（也就是要将全景360度的水平视角和180度的垂直视角映射

[6]这里需要说明的是，垂直模糊仅仅会发生在矩形图块的左右边缘，之所以会出现这种情况，是因为在沿球面的经度进行等距采样时，会致使垂直图块之间出现鬼影效应（具体可参照图 4-**a**）。

[7]HVS的视场角（FOV）相对较大，大约为110° [60]，然而其焦距却仅约为25° [61]。

[8]这里将划分间隔设定为30°，其目的是为了能够覆盖HVS的焦距范围（25°） [61]。

到矩形格式的2D表示形式中）。因此，每个“球面切片（浅蓝色）”都与一个“ERP子图块”相对应；它们在垂直方向上的大小是相同的，但在水平方向上的大小却各不相同。靠近极点的球面切片会在水平方向上被压缩，而这正是导致ERP图像中出现视觉畸变的主要原因。

为了能够让ERP实现无畸变，我们采用了“球面子窗口”，这些窗口具有统一的尺寸（例如图 4-**a**中的红色框所示）。由于HVS的焦距有限，所以这个尺寸应当与最大“球面切片”的尺寸相同——也就是那些位于球面赤道附近的切片。具体来讲，“球面切片”的尺寸范围为（水平：0∼30°，垂直：30°），而“球面子窗口”的尺寸统一为（水平：30°，垂直：30°）。所以，“球面子窗口”通常要比“球面切片”大一些，特别是在靠近极点的区域。通过这样的设置，我们便能够通过运用“球面子窗口”轻松地达成无畸变的投影效果，也就是要将每个“球面子窗口”投影到“ERP子图块”上。

鉴于我们已经获取了“球面切片”与“ERP子图块”之间的映射信息，我们的“网格状球面到2D投影”可以通过以下公式来表示：

$$\begin{aligned}
\text{Sphere} &\leftarrow \mathcal{P}_{\text{E2S}}(\text{ERP}),\\
\{\text{SSlices}, \text{EPats}, \mathbb{M}_{\text{E}\rightleftharpoons\text{S}}\} &= \mathit{SGrid}(\text{Sphere}, \text{ERP}),\\
\{\text{SWindows}, \mathbb{M}_{\text{S}\rightleftharpoons\text{W}}\} &= \mathit{SWindow}(\text{Sphere}, \text{SSlices}),\\
\text{ERP}^{\star} &\leftarrow \mathit{Fill}\big(\text{EPats}, \underbrace{\mathcal{P}_{\text{S2E}}(\text{SWindows})}_{\text{Distortion Free}}, \mathbb{M}_{\text{S}\rightleftharpoons\text{W}}, \mathbb{M}_{\text{E}\rightleftharpoons\text{S}}\big),
\end{aligned} \tag{1}$$

其中“ERP”表示ERP图像，“Sphere”代表ERP的球面表示形式，“SSlices”表示球面切片，“SWindows”表示球面子窗口，“EPats”表示ERP子图块的拓扑结构（具体可参照图 4-**a**）；$\mathbb{M}_{\text{E}\rightleftharpoons\text{S}}$是“ERP子图块”与“球面子切片”之间的关系；$\mathbb{M}_{\text{S}\rightleftharpoons\text{W}}$表示“球面子切片”与“球面子窗口”之间的关系；$\mathcal{P}_{\text{E2S}}$是将ERP投影到球面上的典型投影方式，$\mathcal{P}_{\text{S2E}}$则是将球面投影回ERP的投影方式；$\mathit{SGrid}(\cdot)$用于在球面上执行网格状划分操作，$\mathit{SWindow}(\cdot)$将球面划分为均匀的子窗口；$\mathit{Fill}(\cdot)$以EPats作为指示符，重构无畸变的ERP，也就是最终的ERP*，该过程可以详细表示为：

$$\begin{aligned}
\text{Step 1}: &\quad f_i = \mathit{PAlign}\big(\mathcal{P}_{\text{S2E}}(\text{SWindows}), \mathbb{M}_{\text{S}\rightleftharpoons\text{W}}\big),\\
\text{Step 2}: &\quad f_i = \mathit{PAlign}\big(f_i, \mathbb{M}_{\text{E}\rightleftharpoons\text{S}}\big),\\
\text{Step 3}: &\quad \text{ERP}^{\star} = \mathit{Reform}\big(\underbrace{\{\cdots, f_i, \cdots\}}_{\text{All } f}, \text{EPats}\big),
\end{aligned} \tag{2}$$

其中f是临时容器，$\mathit{PAlign}(\cdot)$依据给定的映射（即$\mathbb{M}$）执行逐像素投影操作，$\mathit{Reform}(\cdot)$将所有无畸变的图块重新组合构建成完整的ERP图；第1步将“无畸变的球面子窗口”投影到“球面切片”，第2步将“球面切片”投影到“ERP子图块”，第3步将得到的“ERP子图块”重新构建为ERP*。关于“球面子窗口”、“球面切片”和“ERP子图块”，请参照图 4-**a**。

关于ERP*的定性展示可在图 4-**a**以及图 4中的②处找到。尽管ERP*的优势在于其所有局部图块均无畸变，但大量的视觉伪影依旧极易被察觉，例如子图块间的错位和鬼影效应（具体可查看图 4中的红色箭头和②）。

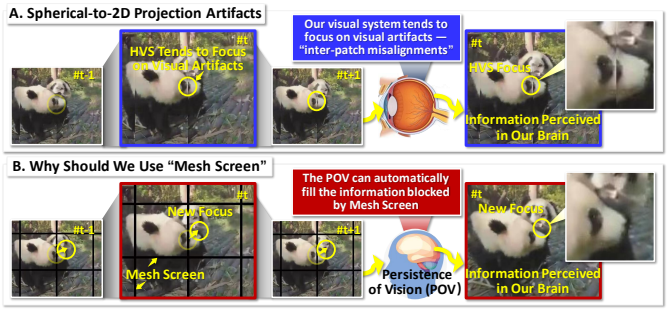

图 5. 为何应使用“网格屏幕”的示意图。子图A和B展示了在视觉输入中有无“网格屏幕”时的视觉信息流。在**子图A**中，由于补丁间存在错位，HVS会持续关注错位区域，而大脑感知到的信息也会集中在补丁间的错位区域。在**子图B**中，补丁间的错位被网格窗口阻挡，这将自动触发大脑的自动填补机制，从而使大脑聚焦于事件本身。详情请见第 3.3节。

3.3 网格屏幕

回顾在第 3.2节所提到的“**前提条件**”，即要让人类视觉系统（HVS）专注于ERP子补丁的内部区域。在本节中，我们提出了一种十分巧妙的解决方案来达成这一前提条件。此外，该方案还能有效减轻ERP*（公式 1）中补丁间的错位问题。具体而言，我们建议在ERP*上应用额外的“网格屏幕”，也就是图 4中③所示的黑色网格屏幕。

我们的技术原理包含两个方面。 **首先**，人类视觉系统（HVS）倾向于关注视觉伪影 [7], [62]（见图 5-A），而采用所提出的“网格屏幕”能够将HVS的焦点转移到每个ERP子补丁的内部内容上（见图 5-B）。由于人类视觉系统通常对规则图案的关注度较低 [47], [63]，而网格屏幕本身属于规则图案，所以能够被自动忽略。再者，我们所设计的网格屏幕的网格大小与ERP子补丁的大小相同，因而可以完全阻挡补丁间的错位情况。所以，网格屏幕能够消除错位现象，并使HVS聚焦于ERP的中心区域。 **其次**，尽管网格屏幕不可避免地会造成一些信息丢失——因为它完全遮挡了补丁间的区域，但我们的大脑能够自动恢复整个ERP的上下文信息。这一现象是通过视野延续（POV）机制 [49], [58], [59]实现的，该机制表明视觉线索在视觉信号消失后仍会在大脑中回响一段时间。由于我们的任务处于视频环境中，所以网格屏幕并不会造成信息丢失，因为POV机制会自动恢复这些区域的信息。

我们的“网格屏幕”由两个部分构成，即网格屏幕生成（*i.e.,* 生成“GMask”）和网格屏幕部署（*i.e.,* 通过⊙），其详细过程如下：

$$\begin{aligned}
&\text{ERP}^{\star\star} \in \mathbb{R}^{w\times h} = \text{ERP}^{\star} \odot \underline{\text{GMask}},\\
&\overbrace{\text{GMask} = \mathit{Grids}(\text{ERP}, \text{EPats}) \in \{0,1\}^{w\times h}}^{\Updownarrow}
\end{aligned} \tag{3}$$

其中，ERP*可通过公式 1获取，ERP**是应用网格屏幕解决方案后的结果；“EPats”在公式 1中已定义——即ERP子补丁之间的拓扑信息，$\mathit{Grids}(\cdot)$提取网格结构，也就是GMask[9]，w，h分别表示ERP图像的宽度和高度；⊙是逐元素乘法。

[9]我们使用“0”表示网格，其厚度为5像素。

借助公式 3，补丁间的错位问题已得到妥善处理（见图 4中的③），当然，“前提条件”也得以达成。接下来，我们将着手解决剩余的“幽灵效应”。

3.4 区分性垂直模糊

尽管我们采用了网格屏幕（第 3.3节）来解决补丁间的错位问题，但我们仍然能够察觉到ERP**（公式 3）中存在的“幽灵效应”，尤其是在靠近极地的那些子补丁中。造成幽灵效应的主要原因在前面已经解释过，即两个水平相邻的“球面子窗口”之间存在重叠区域（见图 4-**a**），参考第 3.2节。

为了处理“幽灵效应”，我们对这些“球面子窗口”的重叠区域进行模糊处理。尽管这个解决方案看起来比较粗糙，可能会导致一些信息丢失，但它却能有效地缓解幽灵效应。其原理与我们提出的“网格屏幕”（第 3.3节）非常相似，即POV机制能够自动帮助大脑恢复这些模糊区域的主要信息。此外，由于运动是吸引注视的一个极为关键的线索，在模糊操作后，这些区域的运动仍有可能吸引HVS的注意[10]。

由于“幽灵效应”在靠近球极的区域更为常见，所以靠近极地的“球面子窗口”将有更大范围的区域需要进行模糊处理，而靠近赤道的子窗口则受影响较小，这就是我们将其称为“区分性垂直模糊（DVB）[11]”的原因，DVB的细节可表示为：

$$\begin{aligned}
\mathrm{SWindows}_{i,j} &\rightarrow \mathrm{Olap}_{i,j} \cup \left\{\mathrm{SWindows}_{i,j} - \mathrm{Olap}_{i,j}\right\},\\
\mathrm{Olap}_{i,j} &= \left\{\mathrm{SWindows}_{i,j} \cap \mathrm{SWindows}_{i,j-1}\right\}\\
&\quad \cup \left\{\mathrm{SWindows}_{i,j} \cap \mathrm{SWindows}_{i,j+1}\right\},\\
\mathrm{SWindows}^{b}_{i,j} &\leftarrow \mathcal{B}\Big(\mathrm{Olap}_{i,j}\Big) \cup \Big\{\mathrm{SWindows}_{i,j} - \mathrm{Olap}_{i,j}\Big\},
\end{aligned} \tag{4}$$

其中，“SWindows”表示球面窗口，已在公式1中定义，首先将其划分为重叠（Olap）区域和非重叠（SWindows-Olap）区域；“Olap”表示重叠区域——即包含幽灵效应的区域；“-”表示减法运算；$\cap$表示交集运算，$\cup$表示并集运算；i和j分别表示球面窗口的行索引和列索引；$\mathcal{B}(\cdot)$是典型的高斯模糊运算[12]；SWindows^{b}是经过我们DVB处理后的输出。

然后，我们的WinDB整体结构可由公式 1变为以下方程：

$$\begin{aligned}
\mathrm{Sphere} &\leftarrow \mathcal{P}_{\mathrm{E2S}}(\mathrm{ERP}),\\
\{\mathrm{SSlices}, \mathrm{EPats}, \mathbb{M}_{\mathrm{E}=\mathrm{S}}\} &= \mathit{SGrid}(\mathrm{Sphere}, \mathrm{ERP}),\\
\{\mathrm{SWindows}, \mathbb{M}_{\mathrm{S}=\mathrm{W}}\} &= \mathit{SWindow}(\mathrm{Sphere}, \mathrm{SSlices}),\\
\mathrm{ERP}^{\star\star b} &\leftarrow \mathrm{Mesh}(\underline{\mathrm{ERP}^{\star b}}),\\
\overbrace{\mathit{Fill}\Big(\mathrm{EPats}, \mathcal{P}_{\mathrm{S2E}}\big(\underline{\mathrm{DVB}}(\mathrm{SWindows})\big), \mathbb{M}_{\mathrm{S}=\mathrm{W}}, \mathbb{M}_{\mathrm{E}=\mathrm{S}}\Big)}^{\text{Discriminative Vertical Blur (Eq. 4)}}&
\end{aligned} \tag{5}$$

其中，大多数符号与公式 1相同；Mesh$(\cdot)$表示提出的网格屏幕（公式 3）；ERP$^{\star\star b}$是公式 5的输出，可在图 4中的④处查看可视化效果。如图所示，靠近赤道的幽灵效应已得到显著缓解，但即便应用了我们提出的DVB，靠近极地的幽灵效应仍然能够被察觉。接下来，我们将对此进一步改进。

图 6. 提出的辅助窗口与动态模糊策略的技术细节。当注视轨迹扫过辅助窗口时，辅助窗口的模糊程度会发生变化（从模糊到清晰）。其优点是能够完全解决幽灵效应，且注视不会被困在辅助窗口中。详见第 3.6节。

3.5 辅助窗口

为了进一步处理ERP$^{\star\star b}$（公式 5）中存在的幽灵效应，我们提出采用“辅助窗口”这种方法。一般来说，辅助窗口的概念并不复杂，就是运用多个定制的“球面子窗口”来辅助用户观看全景场景中的极地区域，可参照图 4中的⑤。设计“辅助窗口”主要依据以下三个原则： **第一**，为了给用户提供更佳的观看体验，所采用的辅助窗口应当在保证无失真的基础上，尽可能拥有较大的覆盖范围。 **第二**，辅助窗口只能占据WinDB主屏幕的一小部分，以此来维持ERP$^{\star\star b}$的“全局”感知特性。 **第三**，辅助窗口应尽可能减少重叠情况，从而有效减轻幽灵效应。

按照这些原则，我们将辅助窗口的覆盖范围设定为垂直45°和水平120°。做出这样的设定主要有以下原因：**1)** 无失真投影$\mathcal{P}_{\mathrm{S2E}}$（在公式 1中定义）的最大水平范围约为120° [42]，而人类视觉系统（HVS）的聚焦范围小于120° [64]，所以我们把辅助窗口的水平覆盖范围确定为120°；**2)** 由于最严重的幽灵效应出现在极地附近，我们依据经验将垂直覆盖范围设定为45°，以便在保持全局感知和抑制幽灵效应之间达到平衡。因此，我们总共设置了N个辅助窗口（见图 4中的⑤），这些辅助窗口是无失真的，放置在上下行用于阻挡幽灵效应，并且大约有70%的ERP$^{\star\star b}$区域仍然可以被访问。

部署辅助窗口的整个过程（DAW）可以详细表示为：

$$\begin{aligned}
\mathrm{WinDB}^{-} &= \mathrm{DAW}(\underline{\mathrm{AWs}}, \mathrm{ERP}^{\star\star \mathrm{b}}),\\
&\overbrace{\mathrm{AW}_i = \mathcal{P}_{\mathrm{S2W}}\big(\underline{\mathrm{SWindows}^{+}}_{i}\big), i \in \{1, ..., N\}}\\
&\overbrace{\mathrm{SWindows}^{+} = \mathit{SWindow}\big(\mathrm{Sphere}, \mathrm{SSlices}^{+}\big),\ \text{Eq. 5}}
\end{aligned} \tag{6}$$

在上述公式中，多数符号已经在公式 5中定义，N表示辅助窗口的总数；与之前的“SWindows”和“SSlices”有所不同，唯一的区别在于SWindows^{+}和SSlices^{+}覆盖了更多的球面区域，即大约是原来的6倍（从$[30°, 30°]$扩展到$[120°, 45°]$）；WinDB^{-}表示我们WinDB的早期版本（见图 4中的⑤），该版本具有多个优点，比如显著减轻了幽灵效应，保持了全局感知。然而，它仍然存在一些缺陷，即由于辅助窗口之间存在“不可避免的重叠[13]”，尤其是注视容易被困在辅助窗口

[10] HVS对运动极为敏感 [63]
[11] 我们保留靠近赤道的两行不变，因为这些区域没有幽灵效应。
[12] 我们根据经验设定高斯模糊的$ksize$=31和σ=5。
[13] 这种重叠主要是由于“‘SWindow^{+}”与“‘SSlices”之间严格的“1对多”关系”导致的。

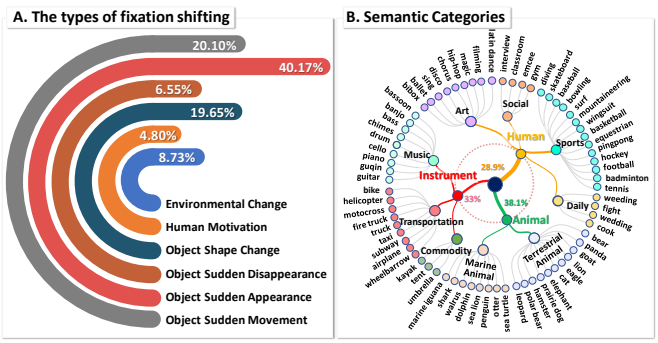

图 7. PanopticVideo-300 数据集中的注视点转移类型和语义类别统计。

中，因为它们相较于较小的ERP子补丁包含更多信息。接下来，我们将解决这些问题，进一步完善我们的WinDB方法。

3.6 动态模糊

在此，我们针对WinDB$^-$（公式 6）中存在的两个问题进行处理：**1)** 少量幽灵效应的残留，以及**2)** 注视被困的问题。如果这两个问题得不到解决，那么所收集到的注视将无法准确反映区域的重要性。因此，我们提出了“动态模糊”方案，其基本思路是对所有辅助窗口进行模糊处理，并动态清除接收到注视的辅助窗口。这样一来，幽灵效应能够被彻底消除，而且注视不会被困在辅助窗口中。此外，由于在模糊的辅助窗口中，运动仍然能够被察觉，所以模糊操作不会导致过多的信息丢失，这一现象在第 3.3节中也已经提及。

我们所提出的动态模糊的技术细节已经在图 6中展示，动态模糊周期性地重复三种状态，即**B**（模糊）、**C**（清晰）和**R**（重新模糊）。 **B**状态：在注视收集开始时，所有辅助窗口均被模糊（采用高斯模糊[14]）。 **C**状态：如果注视轨迹在注视收集过程中扫过某个辅助窗口，该辅助窗口会立即变为清晰。 **R**状态：为防止注视被困在辅助窗口中，辅助窗口的“清晰状态”不会持续太长时间，我们的方法会“逐渐模糊”（持续约2 ～ 3秒）该“清晰”的辅助窗口。我们可以将整个动态模糊（DB）过程表示如下：

$$\begin{aligned} &\mathrm{WinDB} = \mathrm{DB}(\underline{\mathrm{WinDB}^-}), \\ &\overbrace{\{\underline{\mathrm{AW}_i}, i \in \{1, ..., N\}}^{\Uparrow} \\ &\overbrace{\mathbf{B} \xrightarrow{\text{Receive Fixations}} \mathbf{C} \xrightarrow{\text{Last 2s \& No Fixation}} \mathbf{R}}^{\Uparrow} \\ &\quad \uparrow \quad \text{Gradual Blurring, 2}\sim\text{3s} \quad \downarrow \end{aligned} \tag{7}$$

其中 DB(·)表示我们提出的动态模糊方案，该方案作用于WinDB$^-$（公式 6）中的N个辅助窗口； AW_i表示第i个辅助窗口；“WinDB”表示我们提出的新全景注视收集方法的最终版本（见图 4中的⑥）。

总之，我们的WinDB具备以下优点：1）无盲区；2）无幽灵效应；3）无补丁错位；4）良好的全局感知；5）信息丢失最小化；6）用户友好。因此，基于WinDB，我们能够轻松地收集到能够正确反映给定全景场景区域重要性的有效注视。

[14]使用OpenCV GaussianBlur工具，$ksize = 31$和$\sigma = 5$。

这一创新的注视收集工具为全景显著性研究领域奠定了坚实的基础。

4 提出的 PanopticVideo-300 数据集

4.1 为何构建此新数据集？

在以往基于头戴式显示器（HMD）的数据集 [2], [4], [42]里，我们察觉到几乎没有包含“突发事件”的全景视频，而这些突发事件会导致“注视点转移”现象——也就是用户的注视点会从一个地方转移到另一个地方，并且这两处于球面上的距离较远。实际上，“注视点转移”在我们的日常生活中极为常见，并且它可能是一个需要人类视觉系统（HVS）关注的重要情况。然而，由于盲区问题，通过 HMD 收集到的注视数据完全无法察觉这种注视点转移现象。幸亏有我们的 WinDB 方法，现在我们能够处理该问题。因此，基于 WinDB，我们将构建一个全新的数据集，命名为PanopticVideo-300，这是一个极具挑战性且最为全面的全景显著性检测数据集，其中包含了许多复杂场景，并且在这些场景中频繁出现突发事件。

4.2 视频片段收集

为构建上述大型数据集，我们从 YouTube 下载了近 400 个视频片段，其中约 80%的片段包含“突发事件”。随后，我们去除了大概 100 个低质量片段（比如背景单调、动作简单或者分辨率过低的场景）。最终保留了 300 个高质量片段。值得一提的是，先前数据集 [2]–[5]中的视频几乎不存在“突发事件”，这是因为它们在注视收集过程中有缺陷。相比之下，我们的 WinDB 方法即便在“突发事件”出现时，也能够正确收集注视数据。图 7展示了注视点转移类型和语义类别。

4.3 基于 WinDB 的用户注视收集

基于我们新提出的全景注视收集方法（即 WinDB），我们招募了 38 名用户，其中包括 12 名女性和 26 名男性，年龄在 18 至 29 岁之间。所有用户对注视收集过程完全陌生；当然，他们之前未曾看过我们视频片段库中的视频。由于我们的 WinDB 方法无需使用 HMD，所以每位用户仅需在个人电脑（PC）上观看分辨率为 1920×1080 的视频，并且配备典型的眼动追踪仪。每位用户的整个注视收集过程大约需要 50 分钟。倘若用户在注视收集过程中感到疲劳或者不适，可以随时暂停。需要注意的是，这种无需 HMD 的方法（即 WinDB）相较于基于 HMD 的方法更为舒适，更不用说我们的方法收集的注视数据更能符合给定全景场景中区域的重要性程度。

4.4 优势讨论

为彰显我们 WinDB 方法相较于基于 HMD/ERP 的方法的优势，我们在图 8中给出了一些示例，并进行三点深入探讨。

首先，图 8-A 展示了“突发事件”的情形，即重要事件突然发生在与当前主要注视点相距较远的区域。通过对比这

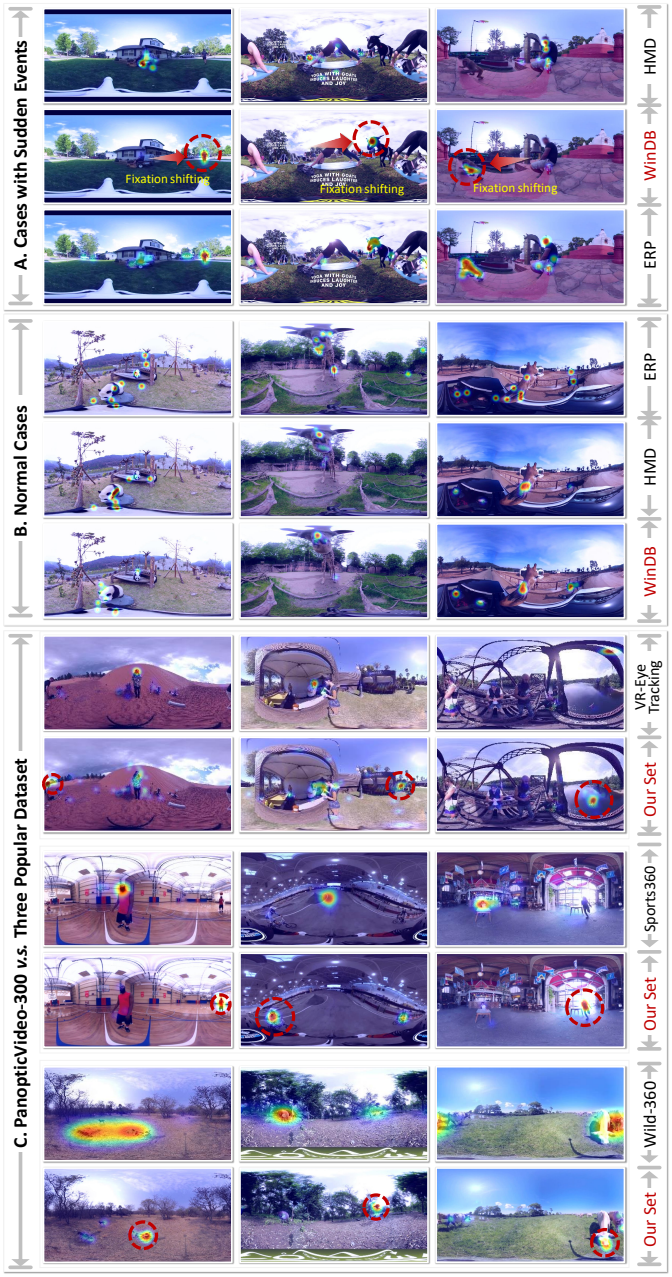

图 8. 我们的方法（WinDB ）与传统注视收集方法（即 ERP 和 HMD）之间的定性比较。注视点转移现象通过红色圆圈突出显示。详见 4.4节。

三行图像，可以发现我们方法收集的注视点（第二行）相较于基于 HMD 的方法（第一行）更为合理。在这些情况下，我们的方法能够捕捉到那些“突发事件”（由红色圆圈高亮显示），而 HMD 方法则无法做到。原因在于这些突发事件发生在 HMD 的盲区范围内，使用 HMD 的用户会错过这些事件。由于我们的 WinDB 方法不存在盲区问题，用户能够充分注意到所有突发事件，从而确保正确的注视收集。

其次，为验证我们 WinDB 方法相较于 ERP 方法的优势，我们在图 8-B 中展示了一些正常情况。在没有“突发事件”的情况下，我们方法收集的注视点与基于 HMD 的方法收集的注视点大体一致，这证明了我们方法的正确性。此外，与基于 ERP 的方法相比，我们 WinDB 方法收集的注视点通常更为集中，而 ERP 方法收集的注视点则较为分散。原因很明显：ERP 中的视觉失真（尤其是靠近极地的区域）容易影响人眼的注视点，使其被失真引发的视觉畸变吸引。由于我们的 WinDB 方法通常不存在失真问题，所收集到的注视点能够更好地聚焦于显著区域。

第三，在上述两个讨论中，我们的团队“重新实现”了竞争对手的方法。为进一步展示我们提出的 PanopticVideo-300 数据集相较于其他真实数据集（即 VR-Eye Tracking [4]，Sports360 [2]和 Wild-360 [42]）的优势，我们在图 8-C 中提供了一些具有代表性的定性比较。我们同样可以得出结论，我们收集的注视点与给定全景场景的真实重要性程度更加契合，原因如前（即无失真且不存在盲区）。

5 提出的FishNet网络

5.1 为何需要此新网络？

当下，我们所拥有的PanopticVideo-300数据库包含了一种独特现象，即“注视点转移”，这一现象使得现有的最先进（SOTA）注视点预测方法 [10], [53], [56]面临着极大的挑战。以下便是引发这一挑战的主要因素。

首先，为了充分利用时空信息（这对于抑制误报极为关键），SOTA方法 [2], [55], [65]在设计网络时设定了保持时空平滑性的要求，以此来约束注视点。然而，“注视点转移”通常呈现间歇性，注视点可能会突然从一处大幅跳跃至另一处（见图 8），这与“正常注视点”（即时空平滑的注视点）有着显著的差异。这两种注视点类型之间的矛盾使得学习过程变得异常艰难。倘若我们直接采用依赖时空平滑约束的现有网络设计，那么注视点转移很可能会被压缩或者被直接忽略。

其次，几乎所有之前的SOTA方法都难以实现全景全局感知，无法察觉突发事件，最终致使它们无法捕捉到注视点转移（如第6.2.3节）。

5.2 问题设定

全景注视点预测的核心目标是学习一个深度模型（NET），该模型以ERP图像作为输入，输出一个概率图，以此表明哪些区域更有可能吸引人类的注意力。给定一个全景图像（即ERP），可通过EF = NET（ERP）来获取预测的注视点（EF），并且NET的学习过程能够借助典型的KL损失来引导 [66]，即KL（EF，GT），其中GT表示之前所收集的真实人眼注视点。问题在于，若采用传统的网络设计（即最简单的Transformer [67]），我们将无法处理注视点转移问题。作为首次尝试，我们提出了一种新颖的网络设计，名为FishNet，用于处理注视点转移，其关键技术创新涵盖**1)** 一种精巧的“全景感知”，用于全局感知突发事件，**2)** 一种全新的“可变形探测器”以及“注视点转移学习”来应对跳跃式注视点。

5.3 网络概述

如图 9所示，我们的FishNet模型主要包含三个主要组件（即“A 全景感知”与“B 可变形探测器”）以及一个定制的学习方案（即“C 注视点转移学习”）。

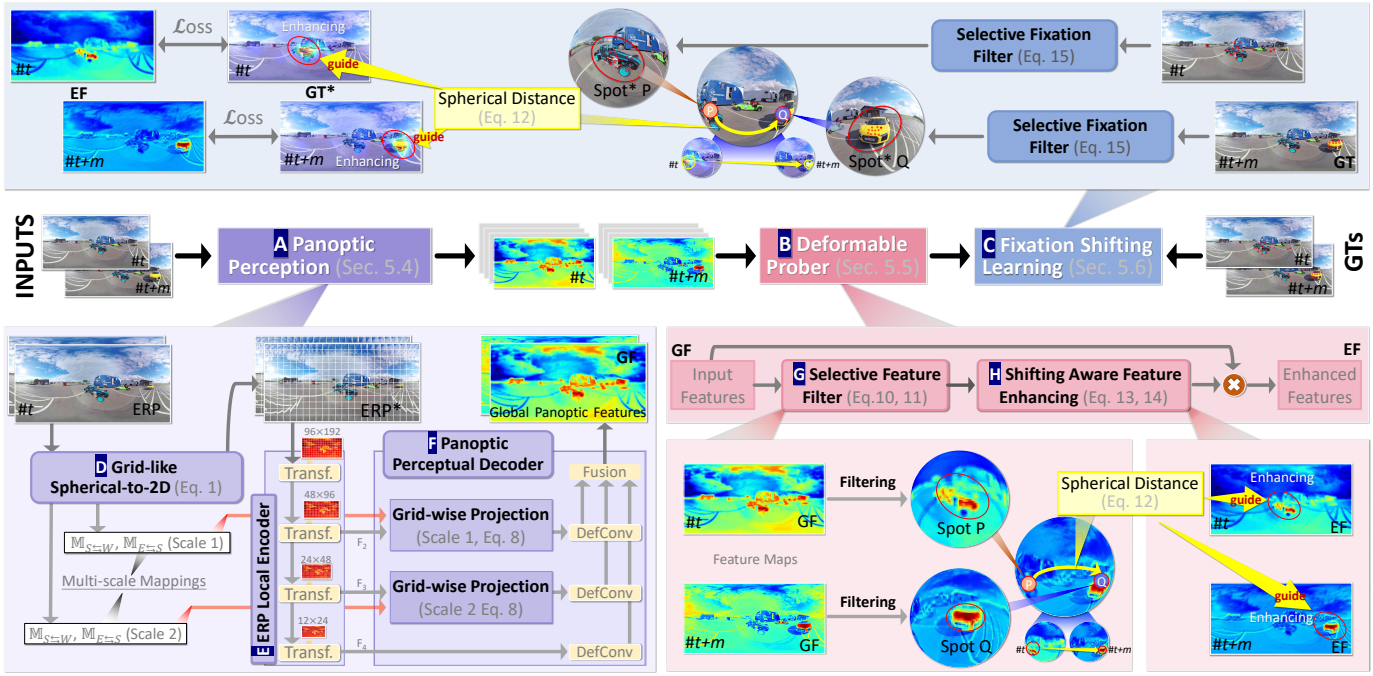

图 9. 我们提出的**Fi**xation **Sh**ifting **Net**work（**FishNet**）的详细网络架构。我们的 FishNet 包含三个主要组件。 **A**部分专注于执行基于 ERP 的全局特征嵌入，以实现全景感知并避免视觉失真。 **B**部分通过重新聚焦网络来捕捉注视点转移，避免了当前最优（SOTA）模型中注视点转移压缩的问题。 **C**部分使网络充分意识到注视点转移机制，确保网络对注视点转移敏感。Transf.：Transformer；DefConv：可变形卷积。有关详细信息，请参见第5.3节。

以ERP图像作为输入，网络能够轻易地捕捉到“突发事件”，因为基于ERP的特征嵌入属于一个全局过程，能够使网络的感知范围覆盖整个全景场景。因此，“A 全景感知”（见图 9）的关键任务是在执行基于ERP的全局特征嵌入时避免视觉失真，具体细节将在第5.4节中详尽阐述。

“B 可变形探测器”的主要目标是使网络能够捕捉到在我们PanopticVideo-300数据集中普遍存在的“注视点转移”。大多数现有的视频注视点预测网络过度依赖“时空”信息，这意味着存在一个强制约束——预测的注视点应保持时空平滑。因此，现有网络更容易将转移的注视点压缩。为了改善这一状况，我们的新型“可变形探测器”使网络能够“重新聚焦”于转移的注视点，且不会产生副作用，具体细节将在第5.5节中介绍。

此外，为了让我们的FishNet对“注视点转移”更为“敏感”，我们还会在训练过程中充分考量给定的GT是否存在注视点转移。若存在注视点转移，训练过程将自动学习其背后的机制。我们借助新设计的“C 注视点转移学习”来达成这一目标，具体细节将在第5.6节中介绍。

5.4 全景感知

5.4.1 技术原理

如图 9 左下角所示，FishNet所提出的“全景感知（Panoptic Perception）”主要由三个部分构成，即 D “网格化球面到二维投影（grid-like Spherical-to-2D）”， E “ERP局部编码器（ERP local encoder）”， F “全景感知解码器（panoptic perceptual decoder）”。

全景感知的技术原理是实现无失真的全局特征嵌入。因此，我们首先运用“网格化球面到二维投影（grid-like Spherical-to-2D）”方法（详见 3.2）将典型的ERP转换为无失真的版本，即ERP*（公式 1）。由于ERP*中的所有子块均无失真，我们将其作为“ERP局部编码器”的独立输入，该编码器是一个典型的基于Transformer的多层次编码器，用于构建子块间的关系。如此一来，“ERP局部编码器”生成的特征便是无失真且具备全局感知能力的。然而，这些特征与包含冗余信息的原始ERP并未能很好地对齐，这些冗余信息主要是由前面在 3.4中提及的“鬼影效应（ghost effects）”所引发的。基于此，我们设计了“全景感知解码器”，通过“网格化投影（grid-wise projection）”将编码器的特征重新投影回ERP结构。如图 9所示，我们进行了两次“网格化球面到二维”投影，以跨越不同尺度的“网格”空间（即Scale 1和Scale 2），目的在于与后续的“ERP局部编码器”多层结构保持一致。经过这一过程后，我们能够获取两个重要的映射信息，即 $\mathbb{M}_{\mathrm{S}\rightleftharpoons\mathrm{W}}$ 和 $\mathbb{M}_{\mathrm{E}\rightleftharpoons\mathrm{S}}$，它们将作为指示器来引导“网格化投影”部分中的特征对齐。

5.4.2 技术细节

“ERP 局部编码器”的特征计算流程可概括如下：

$$\begin{aligned}\{\mathbf{F}_{1,2,3,4}\} &= \overbrace{\mathrm{ERPEnC}\big(\mathit{Split}(\underbrace{\mathrm{ERP}^{\star}}_{\Uparrow})\big)}^{\text{ERP Local Encoder}},\\ &\underbrace{\overbrace{\mathrm{GS2E}(\mathrm{ERP})}^{\text{Grid-like Spherical-to-2D (Eq. 1)}} \rightarrow \{\mathrm{ERP}^{\star}, \mathbb{M}_{\mathrm{S}\rightleftharpoons\mathrm{W}}, \mathbb{M}_{\mathrm{E}\rightleftharpoons\mathrm{S}}, \mathrm{EPats}\}}\end{aligned} \tag{8}$$

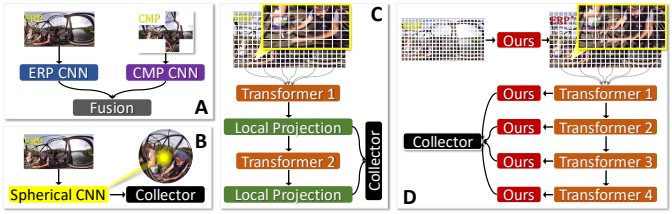

图 10. 我们的方法与现有的全景视频显著性学习方法的技术细节对比。子图 A、B 和 C 表示现有的全景视频显著性学习方法，子图 D 是我们的方法。

在上述公式中，$\mathbf{F}_{1,2,3,4}$代表通过“ERP 局部编码器”所获取的四个中间特征，操作$Split(\cdot)$会把 ERP* 分割成独立的块，以此作为编码器输入（即 ERPEnC）；$\mathbb{M}_{\mathrm{S}\rightleftharpoons\mathrm{W}}$、$\mathbb{M}_{\mathrm{E}\rightleftharpoons\mathrm{S}}$分别是 ERP、球面切片以及球面窗口之间的映射关系；EPats 则是 ERP 网格的拓扑信息，这些均已在公式1里定义，并且会在后续内容（也就是公式9）中被使用。

从图9能够看到，中间特征的尺寸范围是从$\frac{\mathrm{W}}{4}\times\frac{\mathrm{H}}{4}$到$\frac{\mathrm{W}}{32}\times\frac{\mathrm{H}}{32}$，这里的W和H表示 ERP 的大小。由于浅层特征通常较为嘈杂 [68], [69]所提及，所以我们舍弃了$\mathbf{F}_1$。如此一来，实际会使用其中的三个特征（即$\mathbf{F}_{2,3,4}$），这些特征将在我们的“全景感知解码器”中接受处理，目的是将这些无失真且具有全局感知特性的特征表示与原始的 ERP 结构进行对齐。随后，当$\mathbf{F}_{2,3,4}$输入到“全景感知解码器”时，$\mathbf{F}_{2,3}$会被输入“网格化投影（grid-wise projection）”，以此达成上述的特征对齐目标。而未经过“网格化投影”处理的$\mathbf{F}_4$，因其分辨率较低，会直接被当作粗略定位器来使用。因此，“全景感知解码器中的网格化投影（F在图9中）”能够详细描述为：

$$\begin{aligned}\mathrm{GF} = Fusion\big(Concat\big[&\\ \mathrm{DefConv}\big(Fill(\mathrm{EPats}, \mathcal{P}_{\mathrm{E2S}}(\mathbf{F}_2), \mathbb{M}^1_{\mathrm{S}\rightleftharpoons\mathrm{W}}, \mathbb{M}^1_{\mathrm{E}\rightleftharpoons\mathrm{S}})\big),&\\ \mathrm{DefConv}\big(Fill(\mathrm{EPats}, \mathcal{P}_{\mathrm{E2S}}(\mathbf{F}_3), \mathbb{M}^2_{\mathrm{S}\rightleftharpoons\mathrm{W}}, \mathbb{M}^2_{\mathrm{E}\rightleftharpoons\mathrm{S}})\big),&\\ \mathrm{DefConv}(\mathbf{F}_4)\big]\big),&\end{aligned} \tag{9}$$

在这个公式里，GF 是所获取的全局全景特征，“Concat(·)”表示通道级别特征拼接的函数； EPats、$\mathbb{M}$可从公式8得到，$\mathbb{M}$的上标用于表示不同的尺度；函数$Fill(\cdot)$已在公式1中定义，并且在公式2中有详细的说明；$Fusion(\cdot)$是一个典型的特征收集器，包含卷积、批量归一化以及 ReLU 等操作；需要留意的是，“DefConv”表示可变形卷积 [70]，它能够进一步减少（由于“球面子窗口”投影到“球面切片”时存在轻微的覆盖不匹配情况，从而可能产生一些微小的对齐误差）“填充（Fill）(·)”操作之后所产生的微小对齐误差。

5.4.3 全景感知与SOTA解决方案对比

在我们的研究领域中，存在着三种类型的SOTA全景网络；在此，我们将简要阐述我们所具备的优势。

1) 双流网络 [21], [42], [52], [65]。这类网络是由两个研究分支所构成的（见图10-A），具体而言，其中一个分支负责处理存在严重视觉失真的ERP全局信息，而另一个分支则用于处理局部无失真的视图。当将这两个分支结合到一起时，能够使网络的特征同时具备全局感知以及无失真的特性。然而，该网络存在一个关键问题，那就是这两个分支之间的融合难度极大——它们之间并没有明确的对齐方式，这就导致融合过程变成了一个额外的特征嵌入流程。正因如此，采用这种方法所获取到的特征质量相对较差，与我们所得到的结果相比，差距甚远。

2) 基于球面卷积的网络 [2], [10], [14], [45], [53]–[55]。实际上，从理论层面来讲，这类网络的设计是较为完美的，它能够同时实现全局全景感知并且保持无失真的效果（见图10-B）。不过，其关键问题在于所采用的“球面卷积”与现有的CNN存在差异，这就使得所有那些能够提供强大语义特征嵌入的2D特征骨干网络都无法被应用。所以，这些网络在缺乏语义信息支持的情况下，通常表现得并不理想。

3) 基于Transformer的网络 [56], [57]。这类网络在每个Transformer层之后，都要依赖额外的CNN层，也就是所谓的CNN基础的局部投影，以此来处理ERP失真所带来的副作用（见图10-C）。其主要问题在于，所生成的中间特征是通过ERP的补丁来生成的，这些补丁在视觉上与原始的2D特征相比，存在着一定程度的失真，进而导致无法充分利用预训练的特征骨干。

4) 我们的全景感知（Panoptic Perception）方法。如图10-D所示，我们的方法通常是与现有的Transformer相互“独立”的，可以被视作一个通用的插件，其用途在于处理在生成基于ERP的全局特征时所出现的视觉失真问题。需要注意的是，我们的解决方案是支持端到端的训练与测试的。正是由于这种插件的性质，使得网络能够充分利用预训练的特征骨干 [67]。所以，我们的方法所获取到的特征是具备全局属性、无失真特点，并且还具有较强的语义；这些特性使得我们的性能要优于上述所提到的各类SOTA方法。而且，这些特性对于处理“注视点转移（fixation shifting）”现象来说，也是极为必要的，接下来将会对此进行详细的阐述。总之，我们的方法是极具启发性的，所提出的全景感知为特征计算提供了全新的基础，同时也是首次在传统2D研究领域与全景研究领域之间进行搭建桥梁的尝试。

5.5 可变形探测器

5.5.1 技术原理

我们所提出的“可变形探测器（deformable prober）”在图 9 的右下角有相关展示，它主要包含两个部分，也就是“G 选择性特征滤波器（selective feature filter）”以及“H 转移感知特征增强（shifting aware feature enhancing）”。

正如在 5.1 中所提到的那样，学习预测“注视点转移（fixation shifting）”的难度很大，这是因为这些特殊的注视常常会和“正常注视（normal fixations）”产生冲突。具体而言，正常的注视具备时空平滑的特性，而转移的注视却并非如此，它往往会在瞬间从一个地方跳跃到另一个距离较远的地方。由于在实际情况中，“注视点转移”现象相较于正

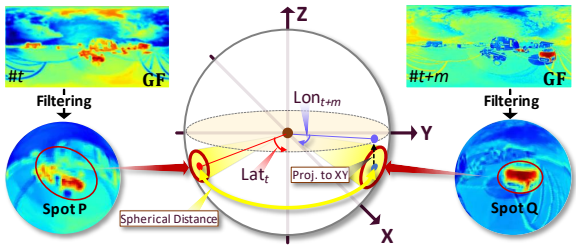

图 11. “**Spot P**”和“**Spot Q**”的球面距离详细计算过程。Spot P和Q来源于FishNet的“选择性特征滤波器（selective feature filter）”部分。其主要目的是测量在较短时间跨度内，分别属于两个不同帧的两个“聚光灯”之间的球面距离。 Lat_t 和 Lon_{t+m} 分别表示第t帧和第$t+m$帧的纬度与经度。详细内容可参见 5.5.2。

常注视要少得多（注视点转移过程在整个视频序列里所占比例较小），所以它很容易被正常注视所掩盖，在网络训练过程中被当作噪声处理，最终导致被忽略掉。因此，我们需要让提出的FishNet网络能够感知“注视点转移”，也就是能够准确地判断是否存在注视点转移以及转移发生的具体位置。所以，“可变形探测器（deformable prober）”的核心原理是利用“选择性特征滤波器（selective feature filter）”来捕捉发生转移的注视，然后借助“转移感知特征增强（shifting aware feature enhancing）”来聚焦于这些转移的注视点。

5.5.2 选择性特征滤波器的技术细节

为了能够捕捉到转移的注视（shifted fixations），我们首先需要明确哪些特征能够将它们与“正常注视（normal fixations）”区分开来。一般来说，转移注视现象具有三个较为特殊的属性。 **其一**，转移的注视应当具有较为强烈的特征响应。 **其二**，与“注视点转移”相关联的区域往往会吸引场景中的大部分注视，也就是所谓的“聚光灯（spotlight）”区域。原因很简单，因为“注视点转移”通常会伴随着“突发事件”的发生；在没有盲区的情况下，用户极有可能被这些区域所吸引并将注意力聚焦于此。 **其三**，在相邻的时域帧之间，这些“聚光灯”区域应当具有较大的球面距离。针对这些属性，我们设计了“G选择性特征滤波器（selective feature filter）”，其过程可以用以下公式表示：

$$\underset{Ms\{\cdots\}\rightarrow \text{Spot}}{\underset{\Downarrow}{\{\text{Fo}_1,\ldots,\text{Fo}_u\}}} = \mathcal{CA}\left(\left|\overbrace{\mathcal{M}(\text{GF})-\mathcal{T}_d\times max\{\mathcal{M}(\text{GF})\}}^{\text{Dynamic Thresholding}}\right|_+\right), \quad (10)$$

在上述公式中，“$\mathcal{M}(\cdot)$”会返回一个矩阵，该矩阵表示其输入的通道均值；GF表示通过Eq. 9所生成的特征；$max(\cdot)$是常见的最大值函数，而$\mathcal{T}_d$则是预先定义好的硬阈值；通过“动态阈值（dynamic thresholding）”处理，结合前面提到的第一个属性，我们就能够较为轻松地获取那些具有高特征响应的区域，这些区域就有可能是包含转移注视的“聚光灯”区域；$|\cdot|_+$表示仅保留矩阵中的正值；$\mathcal{CA}(\cdot)$是连通组件分析（可参考文献 [71]），它能够返回u个孤立的区域（即Fo）；然后，通过函数$Ms(\cdot)$可以对“聚光灯（spot）”进行定位，其原理与上述第二个属性相关，它会返回具有最大特征响应的孤立区域（也就是该区域的平均值）。

因此，上述过程已经满足了“注视点转移”的前两个属性，接下来我们利用“球面距离（spherical distance）”来满足第三个属性，它主要用于衡量在短时间跨度内分别属于两个不同帧的两个“聚光灯”之间的球面距离。该过程可以用以下公式表示：

$$\omega_t = \Big|\Big|\mathcal{P}_{\text{E2S}}\big(F\big(\underbrace{\text{Spot}_t}_{\Uparrow}\big)\big), \mathcal{P}_{\text{E2S}}\big(F\big(\text{Spot}_{t+m}\big)\big)\Big|\Big|_{\text{S}}, \quad (11)$$
$$\text{Spot}_t = SFF(\text{GF}_t),\ \text{Eq. 10}$$

在这个公式中，GF可以通过Eq. 9获得，$SFF(\cdot)$是“选择性特征滤波器（Eq. 10）”，Spot_t是在第t帧中所获取到的聚光灯；$F(\cdot)$会返回其输入的中心坐标，即$\{\text{Lat}_t, \text{Lon}_t\} = F(\text{Spot}_t)$，这里的Lat和Lon分别表示纬度和经度；$||\cdot||_{\text{S}}$用于测量其输入之间的球面距离，详细公式如下：

$$\begin{aligned}&||\{\text{Lat}_t, \text{Lon}_t\}, \{\text{Lat}_{t+m}, \text{Lon}_{t+m}\}||_{\text{S}} = \\ &\quad \arccos\Big(\sin(\text{Lat}_t)\times\sin(\text{Lat}_{t+m}) + \\ &\qquad \cos(\text{Lat}_t)\times\cos(\text{Lat}_{t+m})\times\cos(\text{Lon}_t - \text{Lon}_{t+m})\Big),\end{aligned} \quad (12)$$

我们在图 11 中给出了关于球面距离计算的详细演示。 Eq. 11 的输出ω能够反映第三个属性的满足程度，即较大的ω值意味着Spot_t更有可能是一个包含转移注视的区域。

5.5.3 注视点转移特征感知增强

如图 9所示，借助提出的“选择性特征滤波器（selective feature filter）”，我们已然知晓全景场景中哪些区域存在“注视点转移（fixation shifting）”现象，也就是Spot（公式 Eq. 10所定义的）。为使网络能够聚焦于这些特定区域，我们提出了“H注视点转移感知特征增强（shifting-aware feature enhancing）”这一概念，其核心思路在于着重突出那些与转移注视相关联的特征。我们的想法颇具巧思，主要包含两个按顺序执行的部分：**1)** “点亮”所有可能涵盖转移注视的特征，随后 **2)** 使这些经过修改的特征具备可训练性。在**PART 1**中，我们运用一种较为粗略的手段来达成“点亮”的过程——即依据ω（公式 Eq. 11）简单地增加特征值。换言之，若ω的值越大，那么“Spot”区域的特征值增加幅度就会越大。这一过程的原理十分直观：这些特征在经过粗略的增加处理后，它们与其他特征相比自然会呈现出更为显著的差异，如此一来网络便能够给予它们更多的关注，因为它们在整体训练损失中占据了更大的比重。然而，像上述这样简单地“点亮”特征存在一个关键问题，即尽管网络已经知晓了转移注视的位置所在，但网络仍有可能是“注视点转移不感知的（fixation shifting-unaware）”。也就是说，从网络的视角来看，这些经过修改的特征与常规特征之间并无本质区别，仅仅是某些特征值较大的区域罢了。网络虽然能够从这些特征中进行学习，但其所学内容与“注视点转移过程（fixation shifting process）”毫无关联。因此，我们期望网络所学习的是“注视点转移的过程”，即聚光灯的焦点从一个球面位置转移至另一个位置的这一动态过程。所以，**PART 2**的核心目标

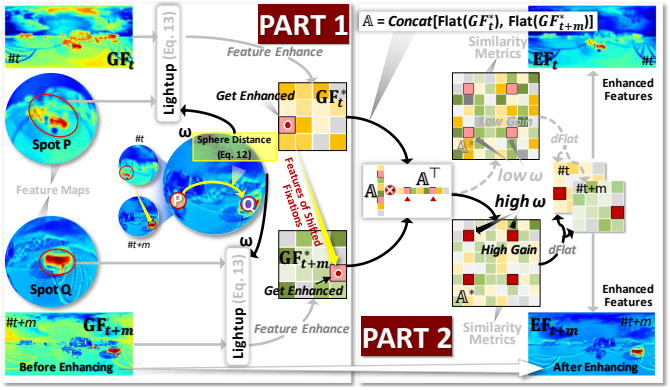

图 12. 可视化“注视点转移感知特征增强（shifting-aware feature enhancing）”包含两个依序执行的部分：**PART 1**，通过增加特征值（即*Lightup*）来强化与转移注视相关的特征，以及 **PART 2**，确保这些修改后的特征能够助力网络理解注视点转移的实际“过程”。详见第 5.5.3节。

便是达成这一期望。接下来，我们将分别对第 1 部分和第 2 部分进行详细阐述，它们在图 12中也有相应的可视化呈现。

GF特征点亮过程（即，**PART 1**）可表示为：

$$\mathrm{GF}^{*} \leftarrow \mathit{Lightup}(\mathrm{GF}, \mathrm{Spot}, \omega), \tag{13}$$

其中，函数 $\mathit{Lightup}(\cdot)$ 依据球面距离ω（公式 Eq. 11）来增加特征值，具体的计算方式为乘以$(1+\omega)$，需要增加的特征由Spot（公式 Eq. 10）所指定。该过程已在图 12的**PART 1**中进行了可视化展示。然后，**PART 2**可表示为：

$$\begin{aligned} \mathit{dFlat}(\underline{\mathbb{A}^{*}}) &\rightarrow \left\{\mathrm{EF}_{t}, \mathrm{EF}_{t+m}\right\}, \\ \mathbb{A}^{*} = \mathbb{A} \odot \sigma\big(\mathit{Softmax}(\mathbb{A} \times \mathbb{A}^{\top}) \times \mathbb{A}\big) & \\ \overbrace{\mathbb{A} = \mathit{Concat}\big(\mathit{Flat}(\mathrm{GF}_{t}^{*}), \mathit{Flat}(\mathrm{GF}_{t+m}^{*})\big)}^{} & \end{aligned} \tag{14}$$

其中，$\mathit{Flat}(\cdot)$ 用于将输入矩阵展平为一个列向量，$\mathit{dFlat}(\cdot)$ 则是将输入重新分割回两个矩阵； $\mathit{Concat}(\cdot)$ 表示连接操作；$\sigma(\cdot)$ 表示sigmoid函数； $\mathbb{A}^{\top}$ 表示矩阵$\mathbb{A}$的转置；EF_{t} 表示第t帧的增强特征。值得留意的是，公式 Eq. 14的关键在于计算$\mathbb{A}^{*}$，它遵循典型的共注意力机制 [72]，将两个不同帧的单独聚光灯合并成“一个统一的注视点转移过程”。整个公式 Eq. 14已在图 12的**PART 2**中进行了可视化展示。

5.6 注视点转移学习

在上述小节中，我们已然提出了一种全新的网络架构，其专门用于解决“注视点转移（fixation shifting）”问题，并且能够借助公式 Eq. 14获取注视点转移感知特征（即，EF）。然而，从网络训练的视角来看，仅仅运用这些注视点转移感知特征并不能确保对良好的注视点转移现象进行有效学习。在此处，我们将进一步阐释这一问题。

通常而言，在视频相关任务（诸如视频显著性 [73]）里，整体的训练损失往往涵盖多个连续的 15 帧片段；而“注视点转移”属于一种突发过程，其可能仅仅发生在两帧之间。在这种情形下，包含注视点转移的帧（也就是这两帧）仅仅占据整个损失的一小部分，这就导致在训练过程中容易忽视“注视点转移”过程。

图 13. 技术细节：将PanopticVideo-300划分为“盲区组”和“普通组”。具体细节参见Sec. 6.1.2。

因此，我们提出了“注视点转移学习（fixation shifting learning）”概念，其目的在于将损失反向传播过程聚焦于包含注视点转移的帧之上。其核心思想与“注视点转移感知特征增强（shifting-aware feature enhancing）”保持一致。与作用于特征有所不同的是，我们的注视点转移学习直接对反向传播的训练损失进行修改，即通过放大包含转移注视的帧的损失来强化学习过程。整个“注视点转移学习”过程如图 9顶部所示，其损失函数能够表示为：

$$\begin{aligned} \mathcal{L}oss = \sum_{t} \mathcal{L}_{\mathrm{KL}}(\mathrm{EF}_{t}, \underbrace{\mathrm{GT}_{t}^{*}}_{\Uparrow}) + \lambda \times \sum \mathcal{L}_{\mathrm{MSE}}\left(\omega_{t}, \omega_{t}^{\star}\right), & \\ \overbrace{\mathrm{GT}^{*} \leftarrow \mathit{Lightup}(\mathrm{GT}, \underbrace{\mathrm{Spot}^{*}}_{\Uparrow}, \omega^{\star}),\ \text{Eq. 13}}^{} & \\ \overbrace{\mathrm{Spot}^{*} = MS\{\cdots\}}^{\Uparrow} & \\ \mathit{Clustering}(\mathrm{GT}) \rightarrow \left\{\overbrace{\mathrm{Fo}_{1}, \ldots, \mathrm{Fo}_{u}}^{\Uparrow}\right\} & \end{aligned} \tag{15}$$

其中，GT 表示真实注视，$\mathcal{L}_{\mathrm{KL}}$ 代表 KL 散度损失 [66]，$\mathcal{L}_{\mathrm{MSE}}$ 表示均方误差损失 [55]，EF_{t} 能够通过公式 Eq. 14获取，ω 和 $\omega^{\star}$ 分别表示特征与真实注视（GT）的球面距离； {Fo} 表示借助 *Clustering*（例如，经典的 DBSCAN [74]）生成的注视簇，函数 *MS* 从 {Fo} 中选取一个注视簇，且被选中的簇应当具备最多的注视点。

6 实验

本实验环节主要涵盖两个层面。 **其一**，我们将详尽地阐述WinDB方法在PanopticVideo-300数据集上所开展的实验情况。通过与借助HMD收集的注视数据展开对比实验，以此来实验验证我们所提出方法的有效性。此外，我们还实施了用户研究，并提供了通用数据集分析，重点在于凸显WinDB和PanopticVideo-300所产生的影响。 **其二**，我们会展示FishNet在PanopticVideo-300上所获取的基准结果，其中包含与当前最优（SOTA）方法进行的定量及定性比较，并且借助消融实验来验证不同组件的有效性。

6.1 有关我们的方法（WinDB）和数据集的实验

6.1.1 平台和软件

我们所采用的WinDB借助Tobii Eye Tracker（v2）来收集全景注视数据。为了有效验证WinDB的有效性，我们还运用HTC VIVE PRO EYE结合7-Invensun-Glass眼动追踪器收集了参考注视数据。在训练与测试代表性模型时，使用了一台配备NVIDIA RTX 3090 GPU的计算机。

6.1.2 数据集划分

鉴于我们的数据集涵盖了“盲区”和“普通”场景，为了便于在后续实验中的“用户研究”以及“通用分析”顺利开展，

表 1
为了验证WinDB中所采用的所有部分的有效性，我们在图 4中进行了组件定量评估。具体细节见Sec. 6.1.3。

O	A	B	C	D	E	CC	SIM	NSS	AUC-J
✓						.297	.293	1.504	.566
	✓					.302	.296	1.531	.568
	✓	✓				.305	.300	1.570	.584
	✓	✓	✓			.306	.301	1.586	.580
	✓	✓	✓	✓		.326	.317	1.592	.588
	✓	✓	✓	✓	✓	**.337**	**.329**	**1.668**	**.596**

O ERP-based Fixations
A Grid-like Spherical-to-2D Projection
B + Mesh Screen
C + Discriminative Vertical Blur
D + Auxiliary Windows
E + Dynamic Blurring

我们将PanopticVideo-300中的视频片段划分成两个组别： **1)** “盲区组”（包含注视点转移的片段）以及**2)** “普通组”（不包含注视点转移的片段）。我们通过测量每15帧中的最大注视点转移距离[15]来判定片段是否包含注视点转移。倘若最大注视点转移距离低于预先设定的阈值（依据HVS的最大视场角设定为110° [1], [75]），那么该片段将被视作“盲区”[16]；反之，则标记为普通片段。这一划分过程如图 13所示。据此，我们的300个片段能够被划分为195个“盲区”片段以及105个“普通”片段。

6.1.3 我们的*WinDB*方法的正确性

我们所提出的WinDB方法相较于经典的基于HMD的方法，其主要优势在于能够妥善处理盲区的限制。在普通的全景场景之中（即不存在盲区的场景），我们方法所收集到的注视数据理应与HMD方法的注视数据相匹配，从而有力地验证了WinDB的可靠性。

为了验证WinDB中各个步骤的有效性（图 4中的A - E），我们开展了组件评估工作，针对每个组件安排了10名用户进行无重叠验证。通过四个指标对用户的注视数据实施定量测试：AUC-J [76]、SIM [77]、CC [78]以及NSS [79]。基于ERP和基于HMD的注视数据均为我们新收集的参考数据[17]。我们从“普通组”（无盲区）的10个片段中随机选取进行验证。

定量结果如表 1所示，充分证明了我们的WinDB（标记为E）显著优于基于ERP的方法（标记为O）。我们能够观察到，一旦某个关键组件得以应用，整体性能便会得到提升，这也证实了每个组件的不可或缺性。顺带一提，我们可能会留意到某些评分并非特别理想。其主要原因在于基于HMD的注视数据并非完美的“真实值”，即“盲区组”和“普通组”的划分是基于一个经验阈值（110°），这在一定程度上限制了数值评分的准确性。

6.1.4 用户实验

我们开展了用户研究，目的在于验证通过WinDB方法所收集的注视数据质量相较于基于HMD的方法究竟如何。实验设置

[15] 由于人类视觉系统（HVS）的响应极限为500毫秒，即约15帧。

[16] 我们为每个15帧片段计算任意两帧的球面距离，并获得一个15×15对角球面距离矩阵，最大元素作为该片段的最大注视点转移距离。盲区组中的每个片段必须至少包含一个具有注视点转移的片段。

[17] 由于在现有的HMD数据集中，注视点转移现象极为罕见，因此我们在全新的PanopticVideo-300数据集上执行基于HMD的注视收集操作，以便进行定量验证。

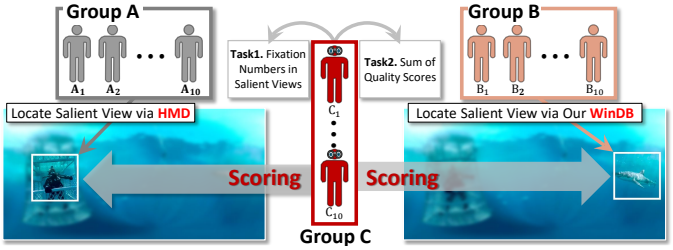

图 14. 主观用户研究的详细信息。用户研究分为三组（即A、B和C），每组各有10名用户。 A组借助HMD收集注视数据并定位显著视图。 B组运用我们的WinDB收集注视数据并定位显著视图。 C组对A组和B组所生成的显著视图进行评分（任务1）。在任务1的评分过程中，C组的注视数据被收集，不过未获取他们的知情同意（任务2）。具体细节见Sec. 6.1.4。

情况如图 14所示，我们从“盲区组”（有关数据集划分的详细技术细节可参照图 13）中挑选了16个视频片段。

为了推进这项用户研究，我们招募了30名用户，并将他们划分为三组（即图 14中的A、B和C组）。 A组和B组分别负责提供基于HMD的注视数据以及基于WinDB的注视数据。随后，我们向C组展示每个片段的局部显著视图，这些显著视图是依据上述相同方案自动筛选出来的。C组的每位参与者需按照0到9的评分标准对每个片段进行打分评价。每个片段将会向C组的用户展示三次。首次展示的是原始ERP版本，其目的在于帮助用户熟悉整体内容。第二次和第三次展示则分别随机呈现基于HMD和WinDB注视数据所选出的显著视图。与此同时，在展示片段期间（除显著视图外的其余区域均进行模糊处理），我们收集了C组参与者的注视数据，这是因为质量更高的显著视图理应能够吸引更多的注视。该项用户研究提供了两个关键指标：**1)** 主观质量评分以及**2)** 显著视图中的注视点数量。需要注意的是，一个优质的视图能够吸引更多用户的注视并且获得更高的质量评分。如图 15所示，实验结果表明，我们的方法在这两个指标上均显著优于基于HMD的方法，并且由我们方法所确定的显著视图吸引了更多的注视且获得了更高的质量评分，由此充分验证了WinDB方法的优越性。

6.1.5 通用分析

本实验旨在验证通过我们WinDB方法所收集的注视数据是否能够助力现有的全景注视预测模型提升其性能。我们选取了三个极具代表性的模型（SpCNN [80]、SalGAN [81]和SalEMA [82]）来进行对比分析。我们的理论依据

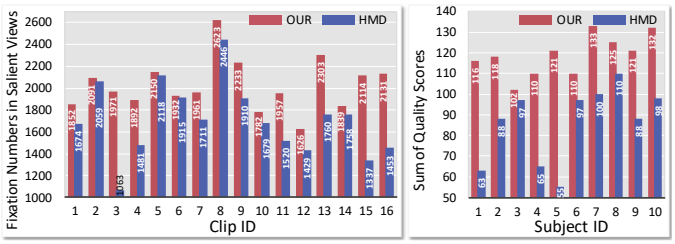

图 15. 用户研究结果。左侧展示了我们方法与基于HMD方法在显著视图中的注视点数量的差异情况。右侧展示了用户在体验我们方法和基于HMD方法后，分别给出的评分总和。这两个结果均表明，我们的方法优于基于HMD的方法。具体细节见Sec. 6.1.4。

表 2

定量评估，验证PanopticVideo-300是否适配现有的基于HMD的注视数据集（即VR-Eye Tracking [4]），具体细节见相关论文内容（**通用分析，Sec. 6.1.5**）。请注意，B1+B2划分表示普通组和盲区组的结合；其对应的趋势可以参见表 3。

Tested on C1									
Dataset	❶ Trained on A1			❷ Trained on A1+A2			❸ Trained on A1+B1		
Model	SpCNN	SalEMA	SalGAN	SpCNN	SalEMA	SalGAN	SpCNN	SalEMA	SalGAN
CC	**.103**	**.130**	**.130**	**.119**	**.153**	**.155**	**.139**	**.159**	**.163**
SIM	.138	.265	.287	.140	.298	.301	.155	.309	.305
NSS	.496	.461	.472	.541	.494	.491	.583	.514	.552
AUC-J	.582	.515	.486	.603	.545	.502	.614	.566	.532
Tested on C2									
Dataset	❹ Trained on A1			❺ Trained on A1+A2			❻ Trained on A1+B2		
Model	SpCNN	SalEMA	SalGAN	SpCNN	SalEMA	SalGAN	SpCNN	SalEMA	SalGAN
CC	**.101**	**.100**	**.109**	**.105**	.134	**.129**	**.156**	**.180**	**.187**
SIM	.134	.258	.256	.136	.274	.285	.182	.319	.330
NSS	.484	.444	.443	.523	.470	.479	.643	.601	.594
AUC-J	.576	.504	.468	.601	.522	.498	.633	.599	.549

A1 50 clips with **ordinary** scenes from **VR-EyeTracking** A2 50 clips with **blind** scenes from **VR-EyeTracking**
B1 50 clips with **ordinary** scenes from **PanopticVideo-300** B2 50 clips with **blind** scenes from **PanopticVideo-300**
C1 30 clips with **ordinary** scenes from **PanopticVideo-300** C2 30 clips with **blind** scenes from **PanopticVideo-300**

表 3

基于PanopticVideo-300、VR-Eye Tracking [4]和Sports360 [2]数据集的FishNet与其他SOTA方法的定量比较。 **TE**、**RT**、**TR**：测试/重新训练/训练模型。 '*'表示方法不加载预训练模型，'+'表示该方法是升级版。

	PanopticVideo-300				VR-Eye Tracking				Sports360					
Methods[year]	CC	SIM	NSS	AUC-J	CC	SIM	NSS	AUC-J	CC	SIM	NSS	AUC-J	Train	Type
ATsalImg$_{att}$ [21]	.113	.145	0.678	.667	.311	.360	1.394	.786	.274	.275	1.585	.824	TE	360
ATsalVideo$_{att}$ [21]	.118	.147	0.767	.679	.338	.374	1.617	.824	.303	.283	1.872	.857	TE	360
GBVS360 [17]	.197	.336	0.727	.689	.348	.396	1.547	.829	.329	.325	1.839	.848	TE	360
BMS [17]	.248	.363	0.867	.720	.353	.397	1.516	.832	.341	.337	1.917	.878	TE	360
ATsalImg [21]	.183	.163	1.042	.757	.242	.330	1.533	.754	.209	.242	1.832	.792	TE	360
BMS360+ [18]	.244	.352	1.006	.738	.361	.402	1.699	.845	.339	.333	1.935	.880	TE	360
GBVS [07]	.196	.168	1.253	.807	.284	.349	1.287	.803	.251	.265	1.384	.806	TE	2D
BMS360 [18]	.259	.372	1.003	.757	.326	.370	1.430	.838	.293	.279	1.624	.863	TE	360
ATsalVid [21]	.200	.165	1.233	.791	.255	.336	1.708	.773	.222	.247	2.045	.807	TE	360
SalGAN [18]	.531	.481	1.692	.823	.418	.401	1.857	.841	.386	.353	2.307	.833	RT	360
SalEMA [19]	.537	.492	1.668	.813	.448	.425	2.236	.841	.473	.382	2.831	.891	RT	2D
*SphCNN [20]	.244	.207	1.096	.734	.271	.347	1.262	.761	.336	.311	1.912	.856	RT	360
TMFI [23]	.503	.473	1.639	.813	.481	.444	2.140	.832	.370	.337	2.004	.878	RT	2D
DATFormer [23]	.550	.509	1.794	.829	.477	.447	2.319	.858	.373	.328	1.932	.828	RT	360
GSGNet [24]	.488	.473	1.610	.813	.434	.414	1.892	.815	.363	.319	1.937	.827	RT	2D
*ADMNet [24]	.451	.419	1.492	.690	.360	.329	1.761	.709	.389	.318	2.449	.651	RT	2D
OUR (FishNet)	**.628**	**.540**	**2.005**	**.853**	**.527**	**.477**	**2.521**	**.874**	**.559**	**.439**	**3.377**	**.926**	**TR**	**360**

主要体现在两个方面： **(1)**在“普通”全景场景中，若通过WinDB收集的注视数据与基于HMD的方法收集的注视数据具有良好的适配性，那么将WinDB收集的注视数据添加到训练集中理应能够显著提升基于HMD注视数据训练的模型的性能表现。 **(2)**通过WinDB收集的注视数据应当能够使全景注视预测模型具备处理“盲区”全景场景的能力。为了验证这两个方面，我们在表 2中开展了相关实验。

我们选定现有的基于HMD的VR-EyeTracking数据集[4]作为基准数据集。从中随机选取50个“普通”场景的片段作为**A1**，50个“盲区”场景的片段作为**A2**。同样地，从我们的PanopticVideo-300数据集中，随机选取50个“普通”场景的片段作为**B1**，50个“盲区”场景的片段作为**B2**。此外，从我们的数据集中，再随机选取30个“普通”场景的片段作为**C1**，30个“盲区”场景的片段作为**C2**。划分出的**A1**、**A2**、**B1**和**B2**将作为三种选定的SOTA模型的训练集，随后在**C1**和**C2**上进行测试。需注意，这些场景的划分之间不存在重叠情况。

通过对比表中的标记❶和❹，我们能够清晰地发现，仅在“普通”场景上进行训练的模型在“盲区”场景上的表现通常较为逊色。这是由于这些模型在仅使用**A1**进行训练时，尚未学会如何处理注视点转移的情形。当使用**A1**+**A2**作为训练集对这三个模型进行测试时，在**C1**和**C2**上的结果呈现出相同的趋势（可参照标记❶ *vs.* ❷和❹ *vs.* ❺），即**C2**的数值通常低于**C1**。然而，我们也留意到，在添加更多的训练数据后，SOTA模型在**C1**和**C2**上的表现均有所改善。这是因为初始的50个片段远远不足以使模型获得理想的性能表现，并且**C2**中的帧并非完全属于注视点转移的情况。最后，通过将**B1**和**B2**添加到训练集**A1**中，我们发现与标记❸和❻相比，模型在“盲区”全景场景上的表现得到了显著提升，这表明**B2**能够有效地帮助模型处理“盲区”全景场景。

综上，该实验表明： **1)**我们的PanopticVideo-300数据集与现有的基于HMD的数据集具有良好的适配性； **2)**在缺乏“盲区”场景的训练数据集上训练的模型难以在存在注视点转移的场景中取得良好的表现——这在现实工作中是极为常见的现象。因此，我们的PanopticVideo-300数据集是对现有全景视频数据集的一项极为重要的补充。

6.2 我们的FishNet网络的实验

6.2.1 *FishNet*的实现细节

我们所构建的FishNet模型采用Transformer [67]作为“ERP本地编码器”（详见Sec. 5.4），并借助PyTorch中的SGD优化器予以实现。将学习率设定为1e-4，批量大小设置为3，总共11个epoch的训练过程。把每个视频帧的尺寸调整为W×H（768×384），并将其划分为lon×lat（即15°×15°）的小块。在训练进程中，每个训练实例涵盖两帧：第t-帧与第$\{t+m\}$-帧，其中m在$\{1, 2, \cdots, 15\}$中随机选取。

6.2.2 定量比较

表 3呈现了我们所提出的FishNet（见图 9）与一些现有的全景注视预测模型在PanopticVideo-300、VR-Eye Tracking [4]以及Sports360 [2]数据集上的性能表现。这些模型包含GBVS360、BMS360算法 [83]、ATsal [65]、SalEMA [82]、SalGAN [81]、SpCNN [80]、SAVT [84]、TMFI [85]、DATFormer [86]、GSGNet [87]以及ADMNet [88]。对于那些具备可用代码的模型，已在我们的数据集上重新进行训练，并在“Train”列中标记为“RT或TR”。正如表 3所示，FishNet在所有其他模型中展现出卓越的性能表现，这凸显了“可变形探测器”与“注视点转移学习”在解决注视点转移问题方面的有效性。此外，由于我们的数据集与现有的基于HMD的数据集具备兼容性（已在Sec. 6.1.5中得到验证），其他模型也能够从性能提升中获益，从而彰显了PanopticVideo-300的普遍优势。

6.2.3 定性比较

图 17展示了FishNet、FishNet$^-$（仅包含“全景感知”组件，排除“可变形探测器”与“注视点转移学习”组件的FishNet版本）与三个具有代表性的SOTA模型（SalEMA [82]、SalGAN [81]以及SpCNN [80]）在PanopticVideo-300上

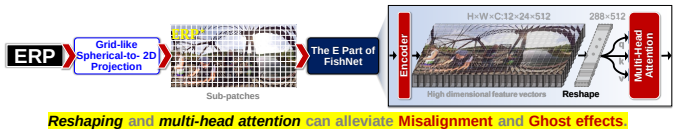

图 16. 关于FishNet为何采用网格状的球面到2D投影作为ERP输入的视觉阐释。详细信息见Sec. 6.2.4。

的视觉效果对比结果。从图中能够看出，FishNet在有效聚焦注视点转移对象方面超越了FishNet$^-$以及其他SOTA模型，这充分展示了“可变形探测器”（详见Sec. 5.5）与“注视点转移学习”（详见Sec. 5.6）的有效性。例如，在图 17的第1行和第3行，FishNet能够精准地聚焦于“猫”和“翼装飞行”的突然出现，而其他模型由于缺乏感知注视点转移的能力，无法将注视点转移至这些突发事件上。此外，凭借其强大的“全局感知”与“局部无失真”能力，FishNet$^-$能够捕捉到诸如第5行中的“黑豹”以及第2行中的“飞行员”等显著事件。

6.2.4 *FishNet*输入：球面到*2D*投影与*WinDB*方法对比

如图 16所示，FishNet完全依托于球面到2D投影方法，这是因为该方法简洁且行之有效，能够应对全景ERP所面临的主要难题——视觉畸变。尽管这种投影方式可能会引发鬼影效应以及错位伪影，但FishNet中的“多头注意力”机制（这是在Transformer中被广泛运用的技术手段）能够高效地处理这些副作用。

从定量比较（表 4）结果来看，WinDB能够带来轻微的性能提升（约 +1%），然而其所需的计算量却高达5.75倍。这一提升，尤其在NSS指标方面，得益于WinDB所引入的较少的视觉伪影。这表明，在处理全景图像固有挑战时，早期进行的显式修改（例如WinDB）相较于后期的隐式注意力机制更为有效。

若要应用WinDB流水线（如图 18所示），可采用双流结构——一个用于ERP**，另一个用于辅助窗口，然后将二者融合。不过，考虑到其高昂的计算成本，我们最终选择了FishNet的球面到2D投影方法，借助多头注意力机制有效地消除视觉伪影。

总之，WinDB旨在用于人类注视数据的收集，其优势在于能够最大程度地减少伪影并妥善解决盲区问题。相比之下，FishNet则专注于通过对全景场景中的兴趣区域进行预测来开展注视学习。由于FishNet的注意力机制自身已具备处理视觉伪影的能力，所以无需像WinDB那样引入较高的计算开销。

6.3 不同组件的有效性评估

6.3.1 *FishNet*中全景感知的有效性

为了验证FishNet中所提出的“全景感知”组件的有效性，我们精心设计了三种不同的实现方式，具体内容如表 5所示。

表 4
不同输入下的定量依据。“WinDB作为FishNet输入”指的是将WinDB作为双分支结构用于注视预测网络，而“FishNet”指的是我们提出的FishNet。

FishNet Input	CC↑	SIM↑	NSS↑	AUC-J↑	FPS↑	ModelSize↓	FLOPs↓
Panoptic Perception	.628	.540	2.005	.852	**11.5**	**97.9M**	**87.02**
WinDB	**.636**	**.543**	**2.022**	**.853**	2.0	241.0M	110.18
	.82%↑	.28%↑	1.70%↑	.05%↑	9.5↓	143.1M↑	23.16↑

表 5
FishNet组件研究的定量依据。（见6.3节）

	O	A	B	C	D	E	F	CC	SIM	NSS	AUC-J
1	✓							.462	.437	1.486	.809
2		✓						.502	.451	1.598	.821
3			✓					**.586**	**.502**	**1.932**	**.836**
4			✓	✓				.594	.507	1.949	.838
5			✓		✓			.598	.515	1.953	.838
6			✓			✓		.611	.520	1.977	.841
7			✓			✓	✓	**.628**	**.540**	**2.005**	**.853**

O	Transformer without Local Projection (No Pre-trained Parameters)
A	Transformer with Local Projection (with Pre-trained Parameters, Fig. 11-C)
B	Panoptic Perception (Sec 5.3)
C	Shifting-aware Feature Enhancing (PART 2 of Fig. 13)
D	Shifting-aware Feature Enhancing (PART 1 of Fig. 13) + Selective Feature Filter (Sec 5.4)
E	Shifting-aware Feature Enhancing (PART 1 and PART 2) + Selective Feature Filter
F	Fixation Shifting Learning (Sec 5.5)

其一，O “无预训练参数”，该方法采用球面卷积网络[2], [53]–[55]，其卷积核定义于球面上，以此实现无失真处理。然而，由于目前并无可用的预训练球面卷积模型参数，所以无法进行加载。因此，为确保公平性，我们所提出的“全景感知”组件在不使用预训练模型的情况下进行实验，并标记为O。

其二，A “带局部投影的Transformer” [56]，该方法在每个Transformer层之后集成了基于CNN的局部投影，以此来处理ERP失真问题。同样地，我们将这些修改应用于所提出的“全景感知”组件中并重新开展训练。

其三，B “全景感知”，这是一种独立于Transformer的方法，能够实现全局全景感知以及局部无失真处理。这种方法能够有效地利用预训练Transformer模型 [67]在2D域的优势。为了与之前的实现方式保持一致（*i.e.*，O和A），我们仅在FishNet中保留“全景感知”组件（*i.e.*，B），并对所有实现方式重新进行训练。

1) 对比表 5中第1行和第3行的数据，我们发现加载预训练参数（*i.e.*，B全景感知）相较于O能够使性能提升3%至10%。这一提升得益于预训练Transformer权重中所嵌入的额外训练数据。

2) 对比表 5中第1行和第2行的数据，我们能够看到，通过利用预训练参数和“局部投影 ($\mathcal{P}_{\mathrm{S2E}}$)”，A带局部投影的Transformer相较于O性能提高了4%。然而，对比第2行和第3行的数据时，我们发现A和B全景感知之间的CC指标存在8%的差距，主要原因在于尽管$\mathcal{P}_{\mathrm{S2E}}$能够减少ERP视觉失真，但它限制了预训练Transformer权重的全面利用。

3) 在表 5的第3行中，我们观察到B全景感知的性能得到了大幅提升。这得益于其独立于Transformer的特性，使其成为一种通用插件。因此，它能够在无需对网络进行修改的情况下实现全局感知，从而能够充分利用预训练的Transformer网络权重 [67]。

6.3.2 可变形探测器的有效性

第 5.5节所阐述的内容中，我们创新性地提出了“可变形探测器”这一概念，其核心目的在于显著强化FishNet针对注视位移现象的学习效能。该方法具备精准捕获并有效强化位移后注视特征的卓越能力，有效避免了这些特征在训练进程中被压缩而沦为噪声，从而保障了特征信息的完整性与有效性。

如图 9清晰所示，“可变形探测器”主要由两个关键构成部分所组成：**1)** “选择性特征滤波器”，其在整个注视位移特征处理流程中发挥着初步筛选与优化的重要作用；以及**2)**

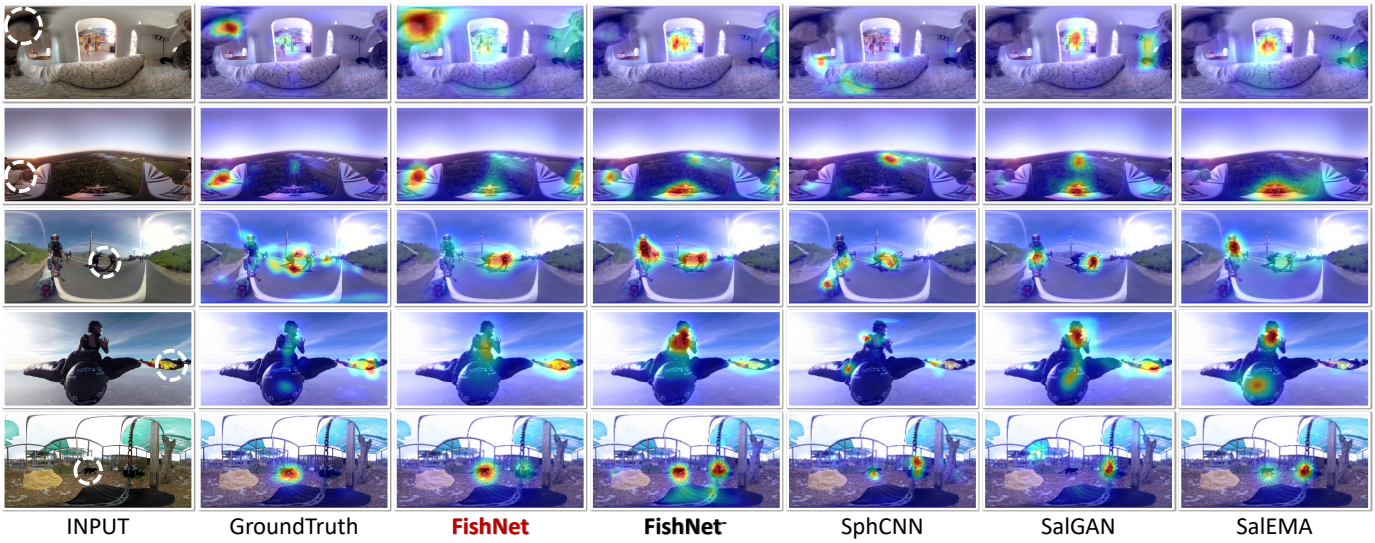

图 17. 我们的模型（*i.e.*，**FishNet**和**FishNet**⁻）与SOTA模型（*i.e.*，SalEMA [82]、SalGAN [81]以及SpCNN [80]）在PanopticVideo-300上的定性比较。**FishNet**⁻指的是仅包含“全景感知”组件的FishNet版本，排除了“可变形探测器”和“注视点转移学习”组件。所有模型均在PanopticVideo-300的训练集上进行了重新训练，并在对应的测试集上进行了测试。详见Sec. 6.2.3。

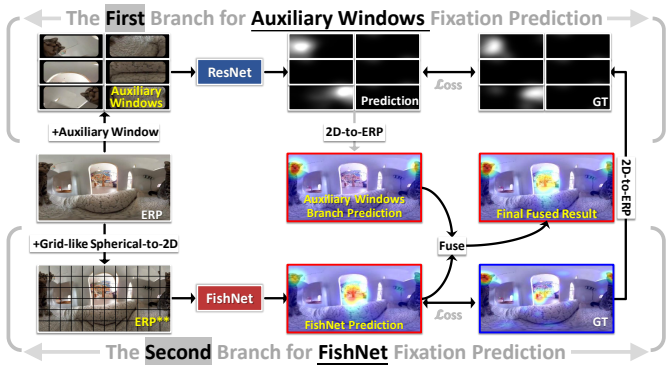

图 18. WinDB作为FishNet输入的双流结构用于注视预测。**第一分支**运用辅助窗口和ResNet进行注视预测。**第二分支**采用FishNet和球面到2D投影处理视觉失真。来自两个分支的结果予以融合以生成最终预测。详细信息见Sec. 6.2.4。

“注视位移感知特征增强”，该部分则聚焦于对已筛选特征进行深度强化与优化处理。

进一步地，图 12将组件 2) 细致地拆解为两个具有明确分工的部分：**Part 1** 着重致力于显著提升那些涵盖位移注视区域的特征值，通过特定的算法与策略，使这些区域的特征在数值层面得到突出体现，从而更易于被网络所识别与学习；**Part 2** 则全力确保这些经过增强处理后的特征具备良好的可训练性，为后续网络能够有效地基于这些特征开展学习与优化工作奠定坚实基础。

依据表 5中的详细数据呈现，当我们对第 3 行和第 4 行进行对比分析时，可以清晰地发现，C 的注视位移感知特征增强（在未包含*Lightup*的情况下）相较于B全景感知，在性能方面实现了约 1% 的提升幅度。

这一现象深刻地表明，即便在缺乏专门针对感知位移注视而精心设计的特定组件时，借助于跨帧相似性矩阵这一强大的工具，注视学习任务依然能够得以顺利开展并实现一定程度的效果，尽管在性能表现上相对而言不够理想，存在较大的提升空间。

当我们将对比的视角转向表中的第 3 行和第 5 行时，能够明显地察觉到，在加入了*Lightup*之后的D选择性特征滤波器，相较于B全景感知，其性能获得了约 2% 的显著提升。

深入探究这一性能提升的内在根源，我们不难发现，这主要得益于*Lightup*这一关键元素能够有效地强化与注视位移紧密相关的特征响应机制。通过这种强化作用，网络对于注视位移的感知能力得到了实质性的提升，从而在整体性能上得以体现出较为明显的进步。然而，需要着重指出的是，尽管这些特征在经过*Lightup*处理后得到了增强，但在缺乏跨帧关联机制的有力支撑下，它们仍然处于相对孤立的状态，难以充分发挥出其最大效能。

最后，当我们对第 3 行和第 6 行进行全面且深入的对比研究时，可以惊喜地发现，我们所精心提出的E可变形探测器（其整合了选择性特征滤波器和注视位移感知特征增强这两大核心组件），相比于B全景感知，在性能层面实现了令人瞩目的 3% 至 4% 的大幅提升。

这一卓越的性能提升成果，主要归因于可变形探测器所具备的强大且全面的功能特性。它不仅能够精准地检测到注视位移现象的发生，还能够对位移后的特征进行深度且有效的增强处理，更为重要的是，它能够在网络层面成功构建起可学习的跨帧特征增强体系，从而使得网络能够充分整合并利用多帧信息，极大地提升了对注视位移的学习与处理能力，最终在整体性能上取得了显著的突破与提升。

6.3.3 注视点转移学习的有效性

在第 5.6节所阐述的内容里，我们创新性地提出了“注视位移学习”这一独特方法。该方法的核心原理在于通过对存在位移注视的帧巧妙施加特定损失，以此来精准且有效地引导整个训练过程，从而促使FishNet网络能够更好地学习和处理注视位移现象。

表 6
对 FishNet的输入大小的消融研究，即，输入 ERP 的大小(Sec. 5.4)。

Input (W×H)	CC	SIM	NSS	AUC-J
1024×512	.560	.478	1.815	.822
896×448	.563	.482	1.859	.828
768×384	**.586**	**.502**	**1.932**	**.836**
640×320	.581	.499	1.898	.833
512×256	.571	.491	1.896	.833

表 7
FisNet中ERP* 子补丁大小的消融研究 (Sec. 5.4)。

Patch (lon×lat)	CC	SIM	NSS	AUC-J
5×5	.574	.494	1.893	.831
10×10	.581	.497	1.885	.833
15×15	**.586**	**.502**	**1.932**	**.836**
30×30	.580	.499	1.892	.831
45×45	.579	.498	1.877	.831

表 8
注视点转移学习中平衡因子的消融研究(Sec. 5.6)。

Loss	λ	CC	SIM	NSS	AUC-J
ω*=0		.611	.520	1.977	.841
$\mathcal{L}_{KLD}+\lambda\times\mathcal{L}_{MSE}$	λ=1	.612	.519	1.982	.842
	λ=3	.614	.522	1.989	.843
	λ=5	**.628**	**.540**	**2.005**	**.853**
	λ=6	.623	.527	1.977	.849
	λ=7	.613	.517	1.967	.847

表 9
选择特征滤波器技术细节中的$\mathcal{T}_d$消融研究(Sec. 5.5.2)。

$\mathcal{T}_d$	CC	SIM	NSS	AUC-J
0.2	.583	.522	1.827	.820
0.4	**.628**	**.540**	**2.005**	**.853**
0.6	.598	.512	1.825	.830
0.8	.580	.512	1.862	.827

从表 5中第 6 行和第 7 行所展示的详细数据可以清晰地看出，F“注视位移学习”相较于E可变形探测器，在性能方面成功提高了 2%。这一显著的性能提升充分证明了“注视位移学习”方法在进一步优化网络对注视位移处理能力方面的重要价值和积极作用。

当我们将目光聚焦于表中的第 3 行和第 7 行并进行深入对比时，能够惊喜地发现，与注视位移相关的组件（即*i.e.*，可变形探测器和注视位移学习）相较于全景感知组件，其性能提升幅度达到了 4%。这一令人瞩目的结果强有力地突显了这些与注视位移紧密相关的组件在处理PanopticVideo-300中复杂的注视位移问题时所展现出的卓越有效性和独特优势。它们相互协同、互为补充，共同构成了一个高效且强大的注视位移处理体系，为提升网络在相关任务中的整体性能做出了极为突出的贡献。

6.4 消融实验

6.4.1 全景感知的输入大小

在第 5.4节所阐述的内容里，我们详细介绍了一种“类网格状的球面到二维（Eq. 1）”转换方法。此方法的核心目标在于有效解决 ERP 失真问题（可参考图 9中的D），通过这一转换过程，ERP 被成功转换为 ERP*，不仅切实有效地解决了视觉失真这一困扰性难题，同时还出色地保持了全局信息的完整性与准确性。

在整个过程中，存在两个“大小”因素会对 FishNet 的性能产生显著影响：其一是输入 ERP 的大小，其二是 ERP*中的子块大小。为了精准确定这两个因素的最佳取值，我们精心策划并开展了两项消融实验，实验所获取的结果分别详细展示在表 6以及表 7之中。

依据表 6中的数据呈现可知，当输入 ERP 的宽度从 512 逐步增大至 768 时，性能实现了 1% 的提升幅度。深入探究这一现象背后的原因，主要是由于更大的 ERP 输入能够更为全面且细致地捕捉到更多的细节信息以及丰富多元的上下文信息，从而为网络的学习与决策提供了更为充足且优质的依据。然而，当进一步将宽度增加到 1024 时，性能却出现了 2% 的下降情况。这是因为随着输入信息的大幅增多，网络处理信息的复杂度也随之急剧增加，进而导致学习过程变得异常艰难，最终对整体性能产生了负面影响。

我们还针对不同的块大小进行了全面且深入的测试，测试结果如表 7所示。经过一系列严谨的实验与分析，我们确定最佳的块大小选择为 15°×15°。当采用其他尺寸时，会致使性能出现略微下降的情况。这主要是因为网络中的相关参数是依据默认尺寸进行精心调整与优化的，当块大小发生改变时，参数与实际输入之间的适配性会受到一定程度的影响。不过，这一微小的性能下降反而从侧面突显了我们所提出方法的良好鲁棒性，即该方法在面对一定程度的参数变化或非最优输入时，依然能够保持相对稳定且可靠的性能表现。

6.4.2 注视点转移学习的平衡因子

在 Eq. 15中，我们创新性地引入了一个因子λ，其核心作用在于巧妙地平衡注视位移损失$\mathcal{L}_{\text{MSE}}$与注视预测损失$\mathcal{L}_{\text{KLD}}$这两个关键要素。通过逐步将λ从 1 增加到 5 的一系列实验过程，我们惊喜地发现性能提升了 2%（具体数据可参考表 8）。然而，当进一步增加λ的值，即从 5 增加到 7 时，却出现了性能下降 3% 的情况。

深入剖析这一现象背后的内在机制，我们可以发现，当λ取值较大时，网络会倾向于过分强调对注视位移的学习，从而可能会忽略其他重要的信息或特征，导致整体性能失衡；而当λ取值较小时，则可能会出现对注视位移学习重视不足的情况，无法充分挖掘和利用与注视位移相关的信息，同样会对性能产生不利影响。经过大量的实验验证与细致的分析比较，我们确定$\lambda = 5$时恰好能够取得最佳的平衡效果，使得网络在注视位移学习与注视预测这两个关键任务之间达到了一种理想的平衡状态，从而实现了性能的最大化提升。

6.4.3 选择性特征滤波器的阈值（$\mathcal{T}_d$）

在第 5.5.2节中的“选择性特征滤波器”部分，我们巧妙地运用了一个动态阈值（$\mathcal{T}_d$）来精准提取“Spot”（即 Eq. 10），这一操作对于准确识别那些可能具有注视位移的区域而言具有至关重要的意义。为了深入探究$\mathcal{T}_d$取值对性能的影响，我们进行了全面的测试，涵盖了不同的$\mathcal{T}_d$值。

表 9中的数据清晰地显示，当增大$\mathcal{T}_d$从 0.2 到 0.4 时，性能能够得到提升（提升幅度为 +2%）。这主要是因为在这一取值范围内，较高的阈值能够更为精准地识别出注视位移区域，从而使得网络能够更加聚焦于这些关键区域进行学习与优化，进而提升了整体性能。

然而，随着$\mathcal{T}_d$的进一步增大，即从 0.4 增加到 0.8 时，性能却出现了 4% 的下降情况。这是因为当阈值过高时，会过度过滤掉一些可能存在位移注视的区域，导致网络获取的信息不完整，无法充分学习到与注视位移相关的所有特征与模式，最终对性能产生了严重的负面影响。

7 局限性

我们所提出的 WinDB 方法在相当程度上有效地解决了头戴式显示器（HMD）注视收集中存在的“盲区”这一棘手问

题。然而，不可忽视的是，WinDB 中的“动态模糊”机制有可能会对用户的注意力产生轻微的偏移影响。例如，在图 19-(a)所展示的情形中，显著事件通常会被定位在黄色辅助窗口内部，当用户持续不断地注视这一显著事件时，便会触发辅助窗口的动态模糊机制。这种情况极有可能会在一定程度上对注视的准确性造成影响。此外，尽管我们所构建的 FishNet 具备感知注视位移的能力，但它在每帧当中仅仅能够聚焦于单一的显著事件。在某些极端特殊的情况下，比如在图 19-(b)所呈现的场景里，一帧之中可能会同时出现多个潜在的“聚光灯”现象，这就极有可能导致对注视位移的判断出现失误。

8 结论

在本文之中，我们成功地提出了三项至关重要的创新成果。其一，我们精准地识别出了在广泛应用的基于 HMD 的全景注视收集方法里所存在的一个关键限制因素，即“盲区”问题，这一问题的存在严重地影响了所收集注视数据的适用性与有效性。

为了切实有效地解决这一问题，我们创新性地引入了一种更为经济实惠、使用起来更加舒适便捷，并且在技术层面上更为精确可靠的无 HMD 方法——**WinDB**。借助这一方法，我们精心创建了一个规模庞大的数据集——**PanopticVideo-300**，该数据集包含了多达 300 个视频片段，其内容广泛地涵盖了 225 个语义类别，其中有 195 个片段涉及到注视位移现象，这无疑为推动全景注视预测领域的进一步发展提供了极为重要的潜力与资源。最后，我们成功地开发出了**FishNet**模型，该模型能够有效地解决注视位移问题，并且通过大量全面且深入的实验充分验证了所有组件的有效性与可靠性，为全景注视预测技术的发展奠定了坚实的基础。

参考文献

[1] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, “Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos,” in *CVPR*, 2017.

[2] Z. Zhang, Y. Xu, J. Yu, and S. Gao, “Saliency detection in 360 videos,” in *ECCV*, 2018, pp. 488–503.

[3] Y. Zhang, F.-Y. Chao, W. Hamidouche, and O. Deforges, “Pav-sod: A new task towards panoramic audiovisual saliency detection,” *TOMM*, 2022.

[4] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao, “Gaze prediction in dynamic 360 immersive videos,” in *CVPR*, 2018.

(a) Salient event kept in poles (b) Multi potential spotlights

图 19. 更多详细信息在Sec. 7，讨论了提议的 WinDB 和 FishNet 的限制。

[5] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, “Predicting head movement in panoramic video: A deep reinforcement learning approach,” *TPAMI*, 2018.

[6] K. Aberman, J. He, Y. Gandelsman, I. Mosseri, D. Jacobs, K. Kohlhoff, Y. Pritch, and M. Rubinstein, “Deep saliency prior for reducing visual distraction,” in *CVPR*, 2022.

[7] L. Jiang, Y. Li, S. Li, M. Xu, S. Lei, Y. Guo, and B. Huang, “Does text attract attention on e-commerce images: A novel saliency prediction dataset and method,” in *CVPR*, 2022.

[8] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, “From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach,” in *CVPR*, 2021.

[9] A. Tsiami, P. Koutras, and P. Maragos, “Stavis: Spatio-temporal audiovisual saliency network,” in *CVPR*, 2020.

[10] Y. Djilali, T. Krishna, K. McGuinness, and N. OConnor, “Rethinking 360deg image visual attention modelling with unsupervised learning.” in *ICCV*, 2021.

[11] Y. Zhu, G. Zhai, Y. Yang, H. Duan, X. Min, and X. Yang, “Viewing behavior supported visual saliency predictor for 360 degree videos,” *TCSVT*, vol. 32, no. 7, pp. 4188–4201, 2021.

[12] H. Yun, S. Lee, and G. Kim, “Panoramic vision transformer for saliency detection in 360 videos,” in *ECCV*, 2018.

[13] A. Nguyen, Z. Yan, and K. Nahrstedt, “Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction,” in *ACM MM*, 2018.

[14] Y. Zhu, G. Zhai, X. Min, and J. Zhou, “The prediction of saliency map for head and eye movements in 360 degree images,” *TMM*, vol. 22, no. 9, pp. 2331–2344, 2019.

[15] C. Jiang, J. Huang, K. Kashinath, P. Marcus, and M. Niessner, “Spherical cnns on unstructured grids,” *ICLR*, 2019.

[16] Y.-C. Su and K. Grauman, “Kernel transformer networks for compact spherical convolution,” in *CVPR*, 2019.

[17] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, “Saliency in vr: How do people explore virtual environments?” *TVCG*, vol. 24, no. 4, pp. 1633–1642, 2018.

[18] D. Chen, C. Qing, X. Lin, M. Ye, X. Xu, and P. Dickinson, “Intra- and inter-reasoning graph convolutional network for saliency prediction on 360 images,” *TCSVT*, 2022.

[19] J. Li, L. Han, C. Zhang, Q. Li, and Z. Liu, “Spherical convolution empowered viewport prediction in 360 video multicast with limited fov feedback,” *TOMM*, 2022.

[20] Y. Lee, J. Jeong, J. Yun, W. Cho, and K.-J. Yoon, “Spherephd: Applying cnns on 360° images with non-euclidean spherical polyhedron representation,” *TPAMI*, 2020.

[21] M. Qiao, M. Xu, Z. Wang, and A. Borji, “Viewport-dependent saliency prediction in 360 video,” *TMM*, vol. 23, pp. 748–760, 2020.

[22] Y.-C. Su and K. Grauman, “Learning spherical convolution for fast features from 360 imagery,” *NeurIPS*, 2017.

[23] Su, Yu-Chuan and Grauman, Kristen, “Making 360 video watchable in 2d: Learning videography for click free viewing,” in *CVPR*, 2017.

[24] M. Xu, L. Jiang, C. Li, Z. Wang, and X. Tao, “Viewport-based cnn: A multi-task approach for assessing 360° video quality,” *TPAMI*, vol. 44, no. 4, pp. 2198–2215, 2020.

[25] D. Tome, P. Peluse, L. Agapito, and H. Badino, “xr-egopose: Egocentric 3d human pose from an hmd camera,” in *ICCV*, 2019.

[26] C. Li, M. Xu, X. Du, and Z. Wang, “Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model,” in *ACM MM*, 2018, pp. 932–940.

[27] H. Kim, H.-T. Lim, S. Lee, and Y. M. Ro, "Vrsa net: Vr sickness assessment considering exceptional motion for 360 vr video," *TIP*, vol. 28, no. 4, pp. 1646–1660, 2018.

[28] A. Erickson, K. Kim, G. Bruder, and G. Welch, "Effects of dark mode graphics on visual acuity and fatigue with virtual reality head-mounted displays," in *VR*, 2020, pp. 434–442.

[29] M. Mahmud, A. Cordova, and J. Quarles, "Visual cues for a steadier you: visual feedback methods improved standing balance in virtual reality for people with balance impairments," *TVCG*, 2023.

[30] C. Sun, M. Sun, and H.-T. Chen, "Hohonet: 360 indoor holistic understanding with latent horizontal features," in *CVPR*, 2021.

[31] J. Xu, J. Zheng, Y. Xu, R. Tang, and S. Gao, "Layout-guided novel view synthesis from a single indoor panorama," in *CVPR*, 2021.

[32] Y. Yang, S. Jin, R. Liu, S. B. Kang, and J. Yu, "Automatic 3d indoor scene modeling fr om single panorama," in *CVPR*, 2018.

[33] C. Zhang, S. Liwicki, W. Smith, and R. Cipolla, "Orientation-aware semantic segmentation on icosahedron spheres," in *ICCV*, 2019.

[34] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2vid: Automatic cinematography for watching 360° videos," in *ACCV*, 2016.

[35] Y. Yu, S. Lee, J. Na, J. Kang, and G. Kim, "A deep ranking model for spatio-temporal highlight detection from a 360° video," in *AAAI*, 2018.

[36] C. Zhuang, Z. Lu, Y. Wang, J. Xiao, and Y. Wang, "Spdet: Edge-aware self-supervised panoramic depth estimation transformer with spherical geometry," *TPAMI*, 2023.

[37] A. Taneja, L. Ballan, and M. Pollefeys, "Geometric change detection in urban environments using images," *TPAMI*, vol. 37, no. 11, pp. 2193–2206, 2015.

[38] L. Jin, Y. Xu, J. Zheng, J. Zhang, R. Tang, S. Xu, J. Yu, and S. Gao, "Geometric structure based and regularized depth estimation from 360 indoor imagery," in *CVPR*, 2020.

[39] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, "Slicenet: Deep dense depth estimation from a single indoor panorama using a slice-based representation," in *CVPR*, 2021.

[40] T. Wu, X. Li, Z. Qi, D. Hu, X. Wang, Y. Shan, and X. Li, "Spherediffusion: Spherical geometry-aware distortion resilient diffusion model," in *AAAI*, vol. 38, no. 6, 2024, pp. 6126–6134.

[41] R. Song, W. Zhang, Y. Zhao, Y. Liu, and P. L. Rosin, "3d visual saliency: an independent perceptual measure or a derivative of 2d image saliency?" *TPAMI*, 2023.

[42] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun, "Cube padding for weakly-supervised saliency prediction in 360 videos," in *CVPR*, 2018.

[43] B. Xiong and K. Grauman, "Snap angle prediction for 360 panoramas," in *ECCV*, 2018.

[44] G. Ma, S. Li, C. Chen, A. Hao, and H. Qin, "Stage-wise salient object detection in 360 omnidirectional image via object-level semantical saliency ranking," *TVCG*, vol. 26, no. 12, pp. 3535–3545, 2020.

[45] Y. Yoon, I. Chung, L. Wang, and K.-J. Yoon, "Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation," in *CVPR*, 2022, pp. 5677–5686.

[46] H. Ai and L. Wang, "Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion," in *CVPR*, 2024, pp. 9926–9935.

[47] Z. Cao, H. Ai, Y.-P. Cao, Y. Shan, X. Qie, and L. Wang, "Omnizoomer: Learning to move and zoom in on sphere at high-resolution," in *ICCV*, 2023, pp. 12 897–12 907.

[48] H. Ai, Z. Cao, Y.-P. Cao, Y. Shan, and L. Wang, "Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions," in *CVPR*, 2023, pp. 13 273–13 282.

[49] H. Ai, Z. Cao, H. Lu, C. Chen, J. Ma, P. Zhou, T.-K. Kim, P. Hui, and L. Wang, "Dream360: Diverse and immersive outdoor virtual scene creation via transformer-based 360° image outpainting," *TVCG*, 2024.

[50] A. Rana, C. Ozcinar, and A. Smolic, "Towards generating ambisonics using audio-visual cue for virtual reality," in *ICASSP*, 2019.

[51] R. Cong, K. Huang, J. Lei, Y. Zhao, Q. Huang, and S. Kwong, "Multi-projection fusion and refinement network for salient object detection in 360° omnidirectional image," *TNNLS*, 2023.

[52] F.-E. Wang, Y.-H. Yeh, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation," *TPAMI*, vol. 45, no. 5, pp. 5448–5460, 2022.

[53] J. Li, L. Han, C. Zhang, Q. Li, and Z. Liu, "Spherical convolution empowered viewport prediction in 360 video multicast with limited fov feedback," *ACM TMCCA*, vol. 19, no. 1, pp. 1–23, 2023.

[54] Y.-C. Su and K. Grauman, "Learning spherical convolution for 360° recognition," *TPAMI*, vol. 44, no. 11, pp. 8371–8386, 2021.

[55] Y. Xu, Z. Zhang, and S. Gao, "Spherical dnns and their applications in 360 images and videos," *TPAMI*, vol. 44, no. 10, pp. 7235–7252, 2021.

[56] H. Yun, S. Lee, and G. Kim, "Panoramic vision transformer for saliency detection in 360° videos," in *ECCV*, 2022, pp. 422–439.

[57] Z. Ling, Z. Xing, X. Zhou, M. Cao, and G. Zhou, "Panoswin: A pano-style swin transformer for panorama understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 755–17 764.

[58] J. Anderson and B. Fisher, "The myth of persistence of vision," *JUFA*, vol. 30, no. 4, pp. 3–8, 1978.

[59] T. Zhang, Y. Shen, G. Zhao, L. Wang, X. Chen, L. Bai, and Y. Zhou, "Swift-eye: Towards anti-blink pupil tracking for precise and robust high-frequency near-eye movement analysis with event cameras," *TVCG*, 2024.

[60] Y. Yoon, I. Chung, L. Wang, and K.-J. Yoon, "Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation," in *CVPR*, 2022.

[61] O. Younis, W. Al-Nuaimy, F. Rowe *et al.*, "A hazard detection and tracking system for people with peripheral vision loss using smart glasses and augmented reality," *IJACSA*, vol. 10, no. 2, 2019.

[62] M. Jiang, L. Shen, M. Hu, P. An, Y. Gu, and F. Ren, "Quantitative measurement of perceptual attributes and artifacts for tone-mapped hdr display," *TIM*, vol. 71, pp. 1–11, 2022.

[63] A. Borst and M. Helmstaedter, "Common circuit design in fly and mammalian motion vision," *Nature neuroscience*, vol. 18, no. 8, pp. 1067–1076, 2015.

[64] W. Yang, Y. Qian, J.-K. Kamarainen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *ICPR*, 2018, pp. 2190–2195.

[65] Y. Dahou, M. Tliba, K. McGuinness, and N. OConnor, "Atsal: An attention based architecture for saliency prediction in 360 videos," in *ICPR*, 2021.

[66] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *CVPR*, 2018.

[67] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[68] W. Gao, S. Fan, G. Li, and W. Lin, "A thorough benchmark and a new model for light field saliency detection," *TPAMI*, 2023.

[69] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "Poolnet+: Exploring the potential of pooling for salient object detection," *TPAMI*, vol. 45, no. 1, pp. 887–904, 2022.

[70] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *ICCV*, 2017, pp. 764–773.

[71] W. Xia, Q. Gao, Q. Wang, X. Gao, C. Ding, and D. Tao, "Tensorized bipartite graph learning for multi-view clustering," *TPAMI*, vol. 45, no. 4, pp. 5187–5202, 2022.

[72] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.

[73] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *TIP*, vol. 27, no. 10, pp. 5142–5154, 2018.

[74] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *AAAI*, vol. 96, no. 34, 1996, pp. 226–231.

[75] S.-H. Chou, Y.-C. Chen, K.-H. Zeng, H.-N. Hu, J. Fu, and M. Sun, "Self-view grounding given a narrated 360 video," in *AAAI*, 2018.

[76] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*.

[77] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *TPAMI*, 2019.

[78] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *TPAMI*, vol. 41, no. 3, pp. 740–757, 2018.

[79] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *VR*, 2005.

[80] D. Martin, A. Serrano, and B. Masia, "Panoramic convolutions for 360 single-image saliency prediction," in *CVPRW*, 2020.

[81] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *ICMEW*, 2018.

[82] P. Linardos, E. Mohedano, J. Nieto, K. McGuinness, X. GiroiNieto, and N. OConnor, "Simple vs complex temporal recurrences for video saliency prediction," in *BMVC*, 2019.

[83] P. Lebreton and A. Raake, "Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images," *SP:IP*, vol. 69, pp. 69–78, 2018.

[84] J. Wu, C. Xia, T. Yu, and J. Li, "View-aware salient object detection for 360° omnidirectional image," *TMM*, pp. 1–15, 2022.

[85] X. Zhou, S. Wu, R. Shi, B. Zheng, S. Wang, H. Yin, J. Zhang, and C. Yan, "Transformer-based multi-scale feature integration network for video saliency prediction," *TCSVT*, 2023.

[86] Y. Zhao, L. Zhao, Q. Yu, L. Sheng, J. Zhang, and D. Xu, "Distortion-aware transformer in 360° salient object detection," in *MM*, 2023, pp. 499–508.

[87] J. Xie, Z. Liu, G. Li, X. Lu, and T. Chen, "Global semantic-guided network for saliency prediction," *KBS*, vol. 284, p. 111279, 2024.

[88] X. Zhou, K. Shen, and Z. Liu, "Admnet: Attention-guided densely multi-scale network for lightweight salient object detection," *TMM*, 2024.

Guotao Wang received his M.S. degree in Computer Science from Qingdao University in 2020. He is currently pursuing a Ph.D. at Beihang University. His research interests include computer vision and deep learning.

Chenglizhao Chen is a Professor in the College of Computer Science and Technology at the China University of Petroleum (East China). His research interests include virtual reality, computer vision, deep learning, data mining, and pattern recognition.

Aimin Hao is a professor in Computer Science School and the Associate Director of State Key Laboratory of Virtual Reality Technology and Systems at Beihang University. His research interests are on virtual reality, computer simulation, computer graphics, and computer vision.

Hong Qin is a Professor in the Department of Computer Science at Stony Brook University. His research interests include geometric and solid modeling, computer graphics, physics-based modeling and simulation, computer-aided geometric design, and scientific computing.

Deng-Ping Fan is a Full Professor and Deputy Director of the Media Computing Lab (MCLab) in the College of Computer Science at Nankai University, China. His research interests span computer vision, machine learning, and medical image analysis.